A Geospatial Analysis of West Nile Virus in the Twin Cities Metropolitan Area of Minnesota

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Debarchana Ghosh

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Dr. Robert B. McMaster and Dr. Steven M. Manson

July 2009

Acknowledgements

Working on a PhD and writing a thesis is certainly a formidable endeavor. However, the last four years have been very enjoyable and a number of people have helped me reach this point.

I would like to thank my advisors, Dr. Robert B. McMaster and Dr. Steven M. Manson, for giving me the freedom to pursue interesting avenues of inquiry but also keeping me on track. The freedom to follow up on different ideas has made the last four years more fun than work. Their advice during these four years has helped me out in various ways such as receiving awards, grant writing, presentation skills, and I have certainly learnt a lot about the process of research from our conversations.

David Nietzel, an epidemiologist from the Minnesota Department of Health and Kirk Johnson, an ecologist from the Minnesota Mosquito Control District played an important role in these last four years. They made it relatively easier to obtain surveillance data of West Nile virus (WNV) infected dead birds, human cases, and infected mosquitoes. In absence of their guidance, I would probably not have learnt as much as I did on the spread of WNV in Minnesota.

I would also like to thank members of the HEGIS and NHGIS laboratories, especially Sula Sarkar, David Van Riper, Petra Noble, Shipeng Sun, Len Kne, and Heather Sanders. Everybody helped me in their own ways ranging from GIS skills to driving a car. We've had a lot enjoyable times both in and out of the lab. Working in these labs was an experience by itself, which provided opportunities to collaborate on various research projects, brainstorm ideas, and help each other out with statistical methods and programming. I am glad that I was part of this group.

I would like to express my gratitude to my parents who were generous enough to accept that I would be far away from them for a long time. In addition, they have always encouraged me to reach higher and this work is a result of their encouragement.

Finally I would like to thank Rajarshi for the support he has given me over the last four years. It has been tough to stay apart from each other for so long, but his support and encouragement became my strength. He has always encouraged me to aim higher and apply for all possible awards, grants, conferences, and publications. In short, without his guidance this thesis would not have been completed.

This dissertation is dedicated to my husband,
Rajarshi

Abstract

The West Nile virus (WNV) is an infectious disease transmitted to humans and other mammals by mosquitoes that acquire the virus by feeding on WNV-infected birds. Since its initial occurrence in New York in 1999, the virus has spread rapidly west and south, causing seasonal epidemics and illness among thousands of birds, animals, and humans. Yet, we only have a rudimentary understanding of how the mosquito-borne virus operates in complex avian-human-environmental systems. The virus first reached Minnesota in 2002 and resulted in several hotspots by 2003. The year 2007 saw one of the severest incidences of WNV in Minnesota. For my dissertation research, I have developed novel approaches to understand the spread and dynamics of the virus by using key environmental, built environment, and anthropogenic risk factors that determine *why*, *when*, and *where* WNV strikes in the Twin Cities Metropolitan area (TCMA).

The first study demonstrates the use of a novel spatiotemporal approach to identify exposure areas. The method retrospectively delineates transmission cycles as exposure areas in their entirety, involving dead birds, mosquito pools, and human cases. Given the strong spatial clustering of WNV infections in the urban areas of TCMA, the next study explores how urban landscape features contributed to the viral activities. This investigation contributed to the broader research question in the field of health geography, of how the heterogeneous urban landscape affects human health and disease patterns. The remaining studies focus on the building and interpreting a nonlinear model which captures the complex relationships between the disease incidences and the hypothesized risk factors. The goal of these studies is to identify risk factor(s) whose management would result in effective disease prevention and containment.

This dissertation has applied contributions to the vector control policies. The findings from the studies can answer two fundamental questions to eliminate larva and adult mosquitoes capable of carrying WNV. First, *when* is the optimal time to apply insecticides and pesticides? Second, *where (area)* should we target spraying of pesticides? This will lead to efficient allocation of resources and allow a balance between mosquito eradication and environmental conservation efforts with respect to insecticide usage.

Table of Contents

List of Tables

List of Figures

1. Chapter 1: Introduction

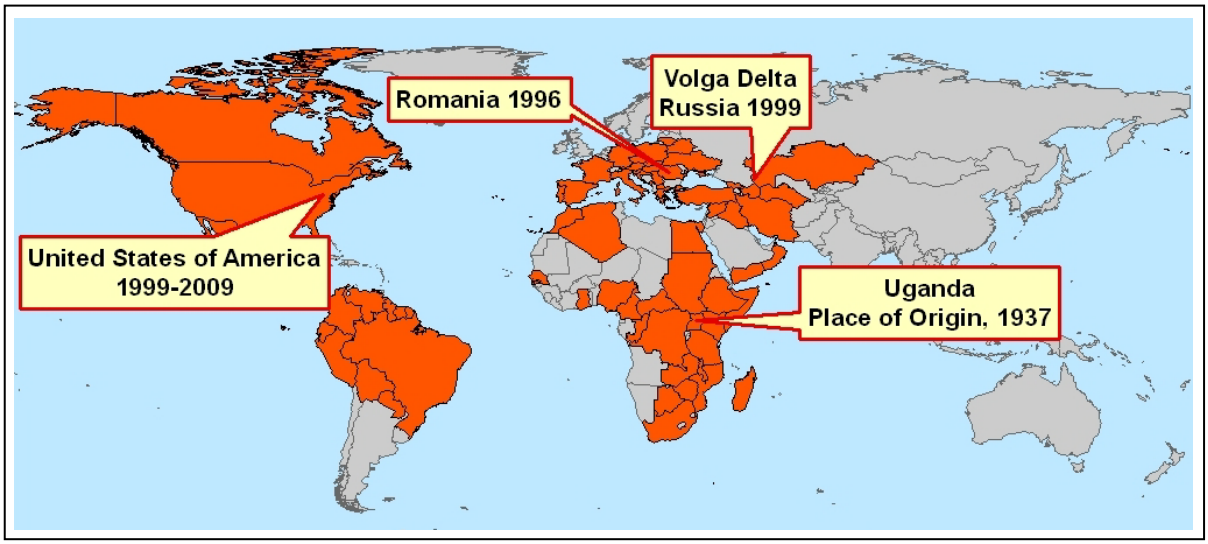
1.1 West Nile Virus

West Nile Virus (WNV), first isolated in Uganda in 1937, is a vector-borne infectious disease of global public health concern. Since its initial occurrence, the geographic spread of this virus has expanded and now includes Africa, Asia, Europe, North America, South America, and the Caribbean (McIntosh et al. 1968; Hubalek and Halouzka 1999; Steele et al. 2000; Malkinson and Banet 2002; OIE 2004; Quirin et al. 2004; Cruz et al. 2005; Matter et al. 2005; Hayes 2007). WNV is transmitted to humans and other mammals by infected mosquitoes that acquire the virus by feeding on WNV-infected birds (CDC 1999). During the previous outbreaks, WNV infection in humans was not considered fatal. Infections in humans varied from asymptomatic symptoms to mild illness with fever, rash, and headache. However during the recent outbreaks there are instances of the more severe form of West Nile, i.e. West Nile encephalitis (inflammation of the brain), and meningitis (inflammation of the lining of the brain and spinal cord), both of which can be fatal.

In the eastern hemisphere, several WNV epidemics resulted in hundreds to thousands cases of infected birds and humans. The human cases were predominantly found among rural populations with fewer diagnosis of severe neurological disorders (Hayes 2007). Interestingly, the recent occurrences of three large WNV epidemics between 1996 and 1999 in southern Romania, the Volga delta in southern Russia, and the northeastern United States (USA) marked the critical changes in the nature of WNV transmission. These changes are as follows. First, the recent epidemics in Europe and

the northeastern USA resulted in hundreds of infected human cases with severe neurological disease and fatal infections. This change was unexpected and therefore became a major public health concern. Second, these were the first epidemics reported in large urban populations unlike the previous ones which were in rural areas. Third, a significant change occurred in the susceptible vector population. In all of the three epidemics, *Culex pipiens*, the common house mosquito, was the principle vector species to amplify and transmit the virus to humans (Hayes 2007). This species had not been previously implicated as the main carrier of WNV. Fourth, the case fatality ratio in birds was very high resulting in significant decline in bird population (Koenig et al. 2007). These changes were critical and put forth WNV as a major emerging vector-borne infectious disease in the country. Therefore prediction, disease pattern, and understanding the risk factors that contributes to the occurrence of WNV incidences are important for minimizing the potential risk for future outbreaks.

Figure 1 The geographic distribution of West Nile virus

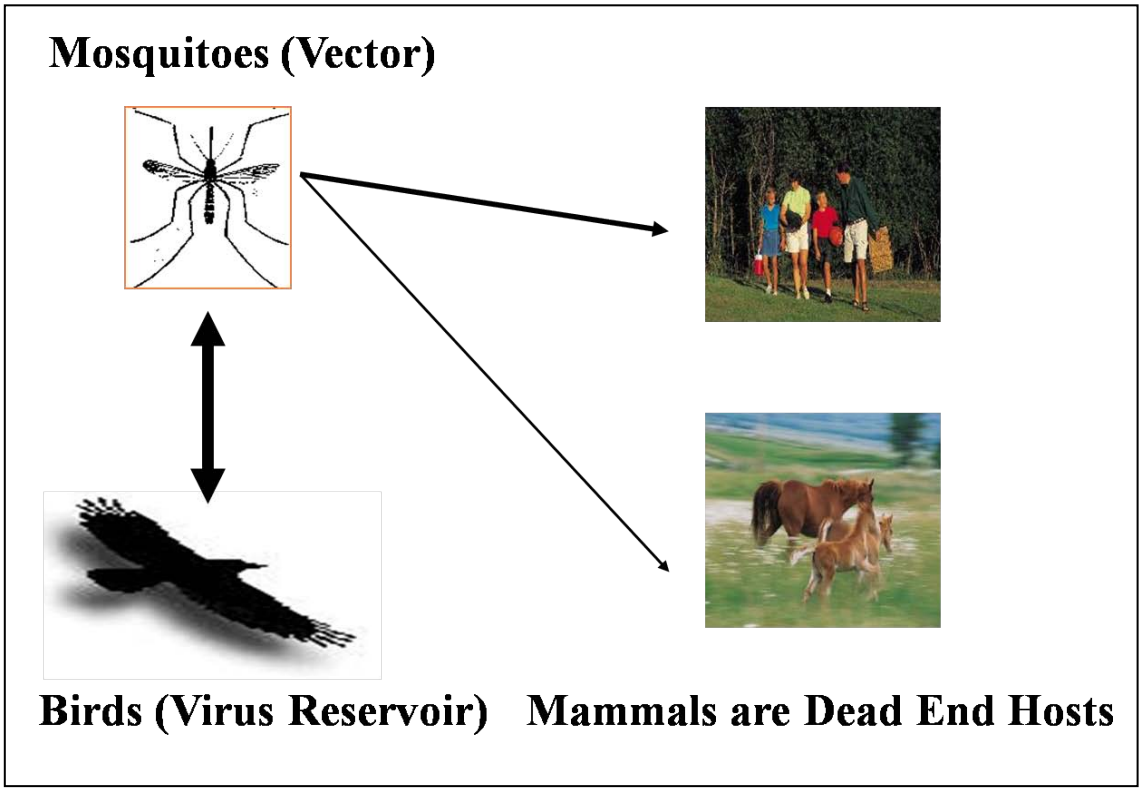


1.1.1 West Nile virus transmission cycle

The three major components in a typical WNV transmission cycle are birds, mosquitoes, and mammals (mainly humans and horses) (Figure 2). Initially, the virus circulates in the blood of birds (reservoir hosts) for a few days after infection. In the Western Hemisphere, especially North America, Blue Jay and American crow are the two common avian species susceptible to WNV infection. Mosquitoes (vectors) become infected when they feed on birds and infected mosquitoes can then transmit the virus to humans and horses through their bites. The virus is also transmitted to other birds when the mosquitoes bite again, thus circulating the virus between hosts and vectors. The virus is injected from the mosquito's salivary glands into its blood stream where it can multiply and cause illness. It was initially believed that direct human-to-human transmission was impossible and that humans are dead end "hosts". However, in 2002, the Centers for Disease Control and Prevention (CDC) discovered the transmission of WNV through blood transfusion and organ transplants as well through breast milk, prenatal infection, and occupational exposure. However human-to-human transmission is very rare.

The incidence of WNV disease is seasonal in the temperate regions of North America and Europe with peak activity from July through October (O'Leary et al. 2004). In the United States the transmission season has expanded due to southward movement of the virus to areas with longer span of warm weather. Therefore climate change and global warming can also affect the length of transmission of WNV.

Figure 2 A typical West Nile virus transmission cycle



1.2 West Nile virus in the United States

In the Western Hemisphere, USA leads with highest number of WNV cases. Since its initial epicenter in New York in 1999, the virus has spread rapidly north, south, and west causing seasonal epidemics and illness among thousands of birds, mosquitoes, humans, and horses. It is also the largest epidemic of human West Nile neuroinvasive disease (WNND) to date in North America (Marfin 2001; Peterson et al. 2003; O'Leary et al. 2004; Huhn et al. 2005). Between 1999 and 2001, few human cases were reported each year; however by 2002, the number increased dramatically to 4,156

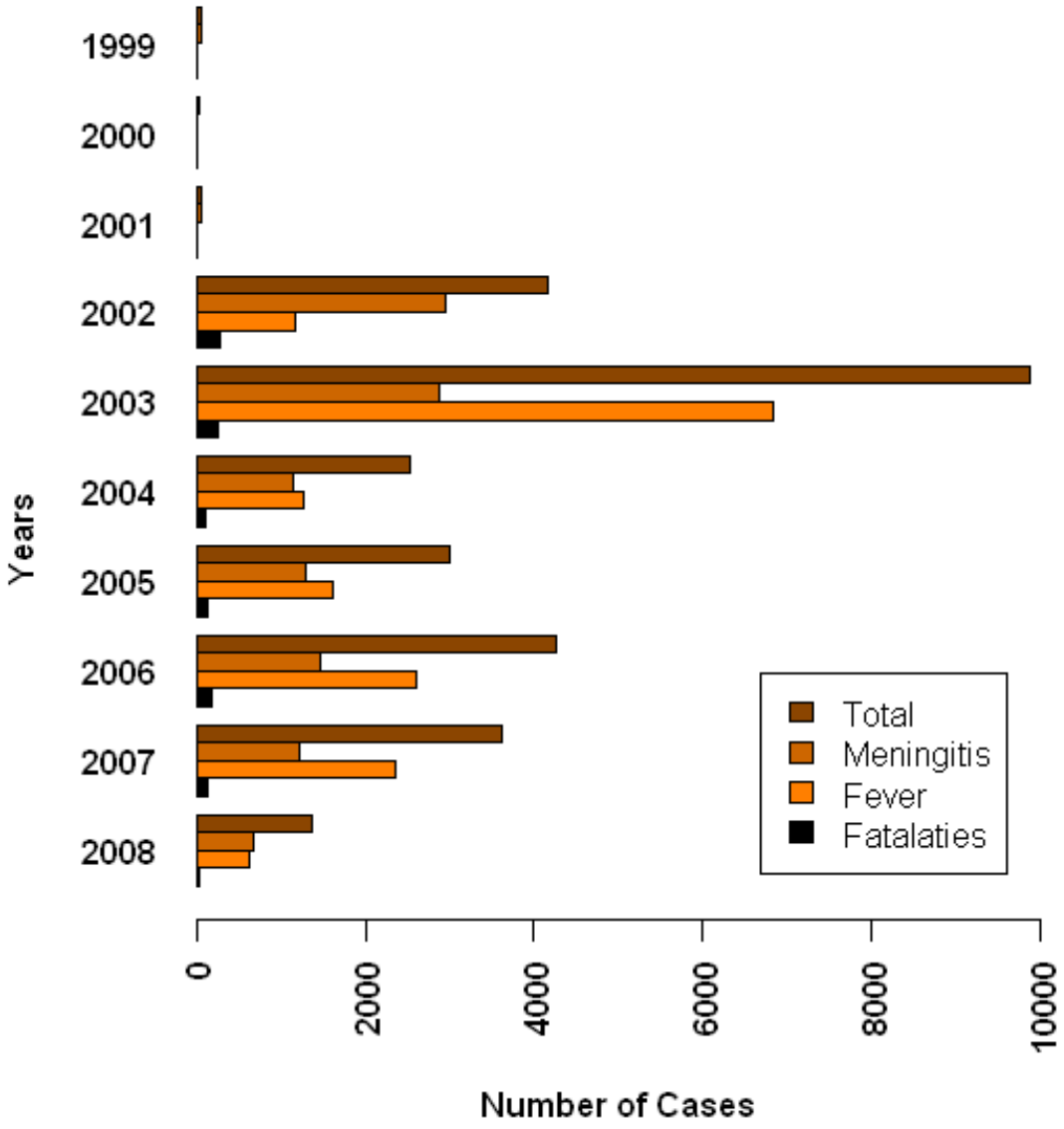
cases. By the year 2004, the numbers increased further to 16,706 human cases, 7,096 of which were classified as WNND, 9,268 as West Nile fever (WNF), and 342 were unspecified clinical cases. Within the next four years there were a total of 28,961 infected human cases, of which 11,753 were diagnosed as WNND, 16,463 as WNF, and 745 were unspecified clinical cases. By the end of 2008, the number of human fatalities due to WNV was 1,131 (Table 1 and Figure 3).

Table 1 Human West Nile virus disease cases by clinical syndrome, United Sates, 1999 – 2008

Year	Total	Encephalitis/ Meningitis	Fever	Other Clinical/ Unspecified	Fatalities
1999	62	59	3	0	7
2000	21	19	2	0	2
2001	66	64	2	0	10
2002	4156	2946	1160	50	284
2003	9862	2866	6830	166	264
2004	2539	1142	1269	128	100
2005	3000	1294	1607	99	119
2006	4269	1459	2616	194	177
2007	3630	1217	2350	63	124
2008	1356	687	624	45	44
Total	**28,961**	**11,753**	**16,463**	**745**	**1,131**

Note: The cases were reported to the Center for Disease Control and Prevention

Figure 3 Human West Nile virus disease cases United Sates, 1999 – 2008



The transmission of WNV has spread dramatically in the United States (Figure 4). From 2001 to 2002, intense transmission shifted from the northeastern states to Midwest, where states like Illinois (884 cases) and Michigan (644 cases) led the nation in the number of human cases. Some of the important focal points with very high local rates were found around Chicago and Detroit. This urban-centric nature of transmission of the virus in the Midwest reflected similar patterns seen with the spread WNV in the eastern United States. By 2003, severity of transmission shifted from the Midwest and south-central states to the western plains and Front Range of the Rocky Mountains. Here, states like Colorado (2947), Nebraska (1942), and South Dakota (1039) reported human cases well above the 1000 threshold. This was also the year with highest number of infected human cases (9862) in the United States (Table 1) to date. In the following couple of years, 2004 and 2005, cases were reported in the western states of California, Arizona, and western Colorado. However there were no significant clusters with high incidence rates. (Figure 4). From 2006 and 2007, the concentration of human cases shifted again to the western plains and upper Midwest including states like Idaho, Colorado, Nebraska, North Dakota, and South Dakota. In summary, the spread of WNV incidences in the United States showed significant spatial trends, especially from east to west with largest outbreaks occurring in 2002, 2003, and 2006. There were also evidences of high local rates of infection in humans around Chicago (Ruiz et al. 2004; Ruiz et al. 2007), Detroit (Ruiz et al. 2007), urban areas in the northeastern U.S. (Brown et al. 2008), and Georgia (Gibbs et al. 2006).

Figure 4 Reported incidence of neuroinvasive West Nile virus disease by county, United States, 1999-2007

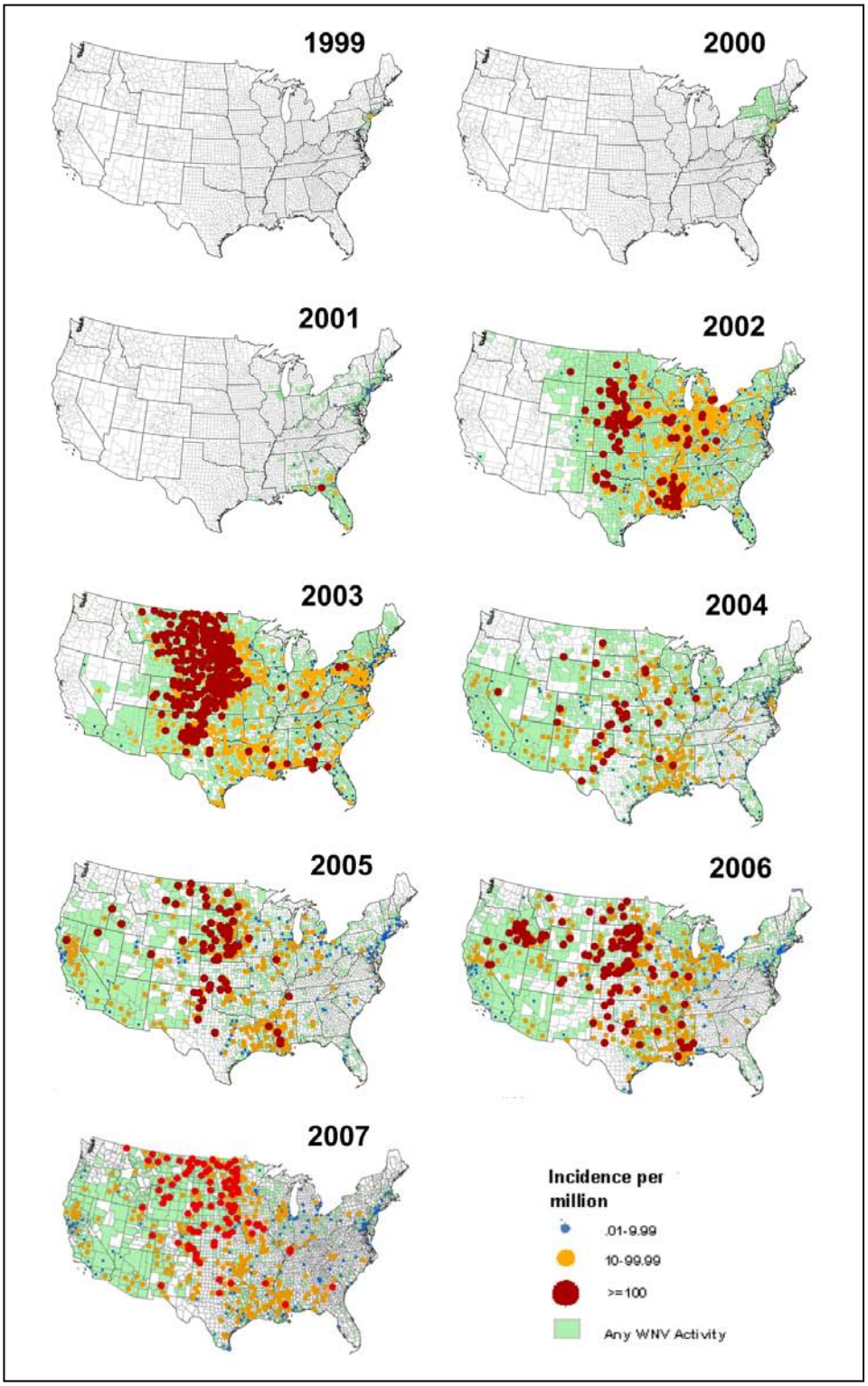


Figure 4. Source: The Center for Disease Control and Prevention (CDC). These West Nile virus maps reflect surveillance reports released by state and local health departments to CDC's ArboNET system for public distribution. The map shows the distribution of human neuroinvasive disease (encephalitis and/or meningitis) incidence with number of human cases shaded according to incidences ranging from 0.01 to 9.99, 10 to 99.99, greater then 100, and WNV activity (human, mosquito, veterinary, avian and sentinel data).

1.3 West Nile virus in the state of Minnesota

The virus first appeared in the state of Minnesota in the year 2002. There were 341 infected dead birds (American crow and Blue jay), 48 human cases, of which 16 were WNND and 32 were diagnosed with WNF (Table 2). In addition, the infection was intense in horses with 997 infected cases. According to the Metropolitan Mosquito Control District of Minnesota (MMCD), the four main vector species potential to carry the virus are *Culex pipeins, Culex restuans, Culex tarsalis,* and *Culex salinarius*. By 2003, WNV infections resulted in several regions of special epidemiological concern or 'hotspots' with a total of 344 dead birds and 152 infected human cases. One such region of public health concern was the urban and suburban areas of Minneapolis and Saint Paul. The occurrence of WNV incidences in birds, mosquitoes, and humans exhibited strong clustering around the Twin Cities of Minnesota. This was also the year with first cases (four) of human deaths due to WNV. However the number of infected equine cases dropped significantly, primarily because of the introduction of vaccines for horses against WNV infection. With lower case counts between 2004 and 2005, the rate of infection intensified again in the year 2006 and 2007. In 2007, human cases increased to 103, with 44 WNND cases, 57 WNF fever symptoms, and 2 deaths (Table 2). Thus from 2002 to 2008, there were a total of 1598 infected dead birds and 473 human cases, with highest intensity of infection in 2003, 2006 and 2007 (Table 2).

Table 2 West Nile virus incidences in Minnesota, 2002 – 2008

Year	Total	Encephalitis/ Meningitis	Fever	Fatalities	Bird	Veterinary
2002	48	16	32	0	341	997
2003	152	48	100	4	433	82
2004	35	12	21	2	159	10
2005	48	18	27	3	91	22
2006	77	31	43	3	500	17
2007	103	44	57	2	65	17
2008	10	2	8	0	9	0
Total	**473**	**171**	**288**	**14**	**1598**	**1145**

Note: The cases were reported to the Center for Disease Control and Prevention, Minnesota Department of Health, and Metropolitan Mosquito Control District of Minnesota.

1.3.1 West Nile virus in the Twin Cities Metropolitan Area

The Twin Cities Metropolitan Area of Minnesota (TCMA) includes seven counties of Anoka, Hennepin, Dakota, Carver, Scott, Ramsey, and Washington (Figure 5). The 2002 WNV epizootic resulted in 102 reports of infected dead birds and five infected mosquito pools in the metropolitan area (Table 3). In addition, 13 human cases and 141 equine cases were reported to the Minnesota Department of Health (MDH). In 2003, clusters of WNV incidences were found in the Twin Cities of Minneapolis and Saint Paul and its surrounding suburban areas with 26 human cases, 285 dead birds, and approximately 1400 infected mosquitoes. This is more evident in the spatial distribution of WNV incidences in 2002 and 2003 (Figure 6).

Figure 5 Location of the Twin Cities Metropolitan Area of Minnesota

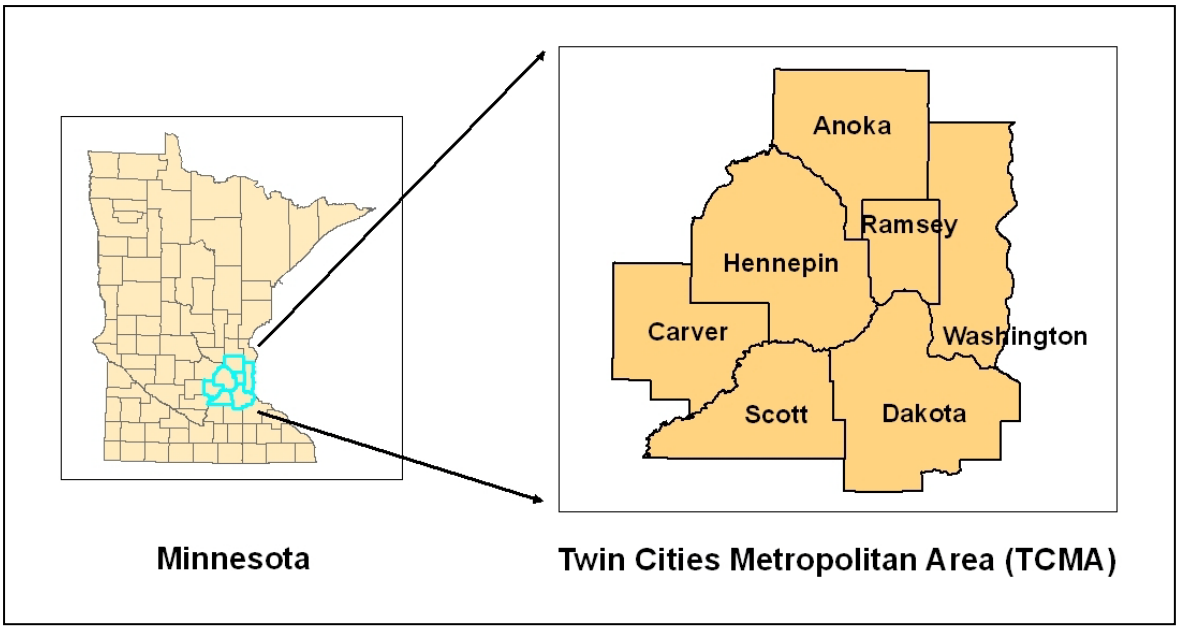


Table 3 West Nile Virus incidences in the Twin Cities Metropolitan Area of Minnesota, 2002 - 2007

Year	Human Cases	Infected Dead Birds	Positive Mosquitoes Pools	No. of Mosquitoes	Veterinary Cases
2002	13	102	5	1100	141
2003	26	285	17	1400	14
2004	6	125	2	25	0
2005	7	60	14	160	5
2006	15	479	90	1550	2
2007	18	60	85	1560	4
Total	**85**	**1111**	**213**	**5795**	**166**

Note: The cases were reported to the Minnesota Department of Health and Metropolitan Mosquito Control District of Minnesota.

Figure 6 Spatial distribution of West Nile Virus incidences in the Twin Cities Metropolitan Area of Minnesota in 2002 (Left figure) and 2003 (Right Figure)

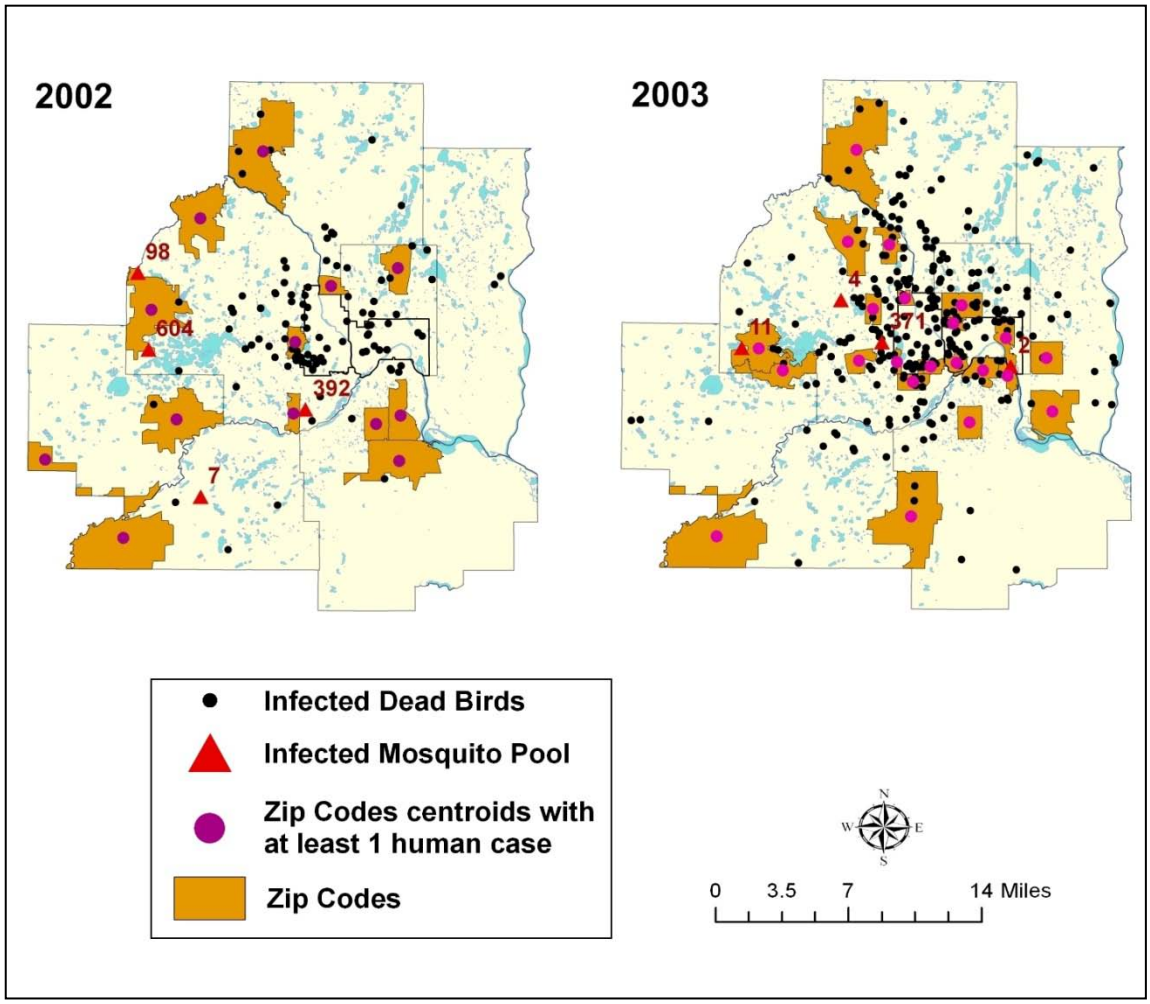


Figure 6: Sources are Minnesota Department of Health and Metropolitan Mosquito Control District. The maps show the distribution of human cases aggregated by zip codes, with their centroids as the points of occurrence, and location of infected dead birds, and positive mosquito pools as points.

In 2004 and 2005 the number of WNV cases declined and were scattered across the TCMA (Figure 7). However in 2006, highest number (479) of WNV infected dead birds were reported mostly clustered around Minneapolis and Saint Paul. Further, there were approximately 1550 infected mosquitoes from 90 WNV-positive mosquito pools, and 15 human cases (Table 3). Even though the number of dead bird reports declined in 2007, a large proportion of vector population was infected, which was successful in transmitting the virus to humans. Similar to the previous years of 2002 and 2003, the spatial distribution of disease occurrence in 2006 and 2007 indicated urban preference. This reflected similar patterns seen with WNV transmission in the eastern United States (Mostashari et al. 2003a; Hayes et al. 2005; Brown et al. 2008), Chicago, Detroit (Ruiz et al. 2007), and Georgia (Gibbs et al. 2006).

Figure 7 Spatial distribution of West Nile Virus incidences in the Twin Cities Metropolitan Area of Minnesota in 2004 (Left figure) and 2005 (Right Figure)

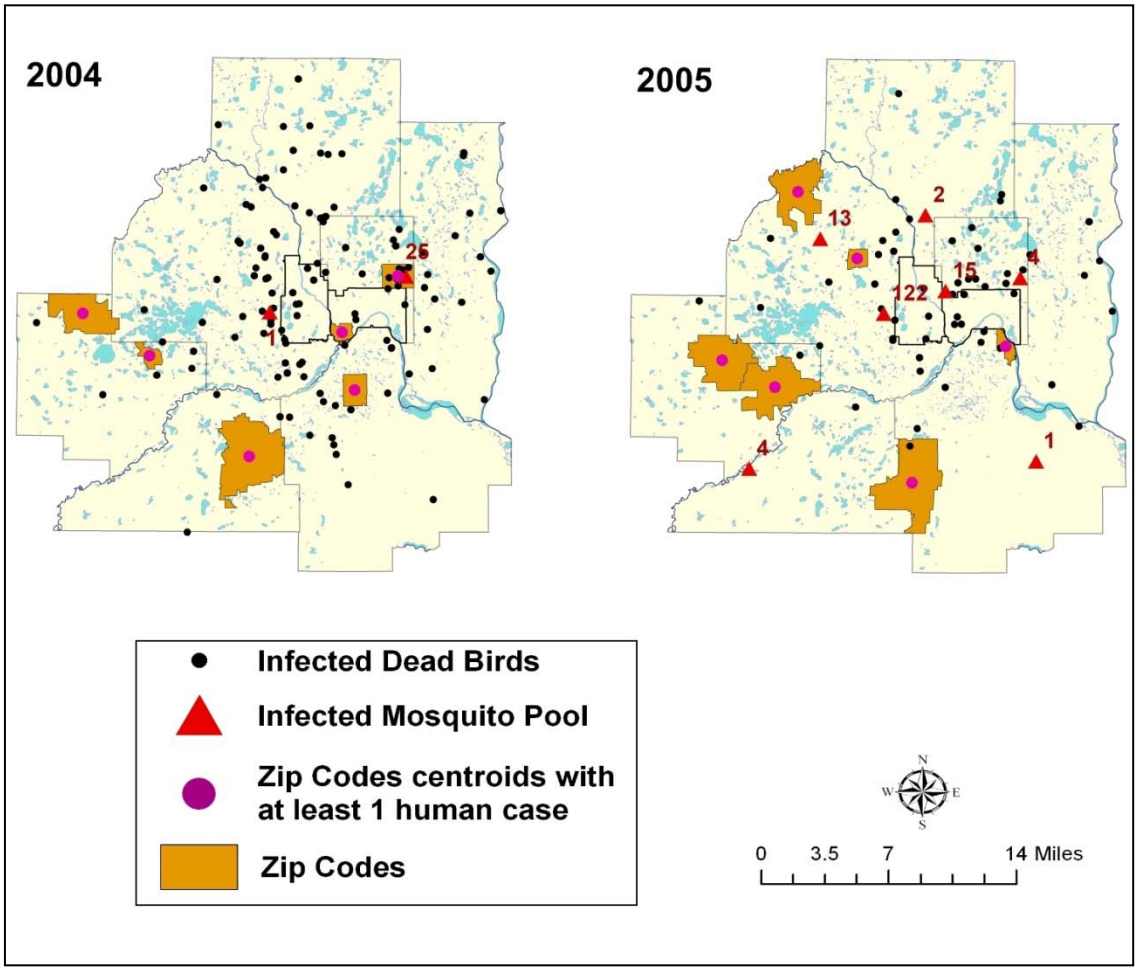


Figure 7 Sources are Minnesota Department of Health and Metropolitan Mosquito Control District. The maps show the distribution of human cases aggregated by zip codes, with their centroids as the points of occurrence, and location of infected dead birds, and positive mosquito pools.

Figure 8 Spatial distribution of West Nile Virus incidences in the Twin Cities Metropolitan Area of Minnesota in 2006 (Left figure) and 2007 (Right Figure)

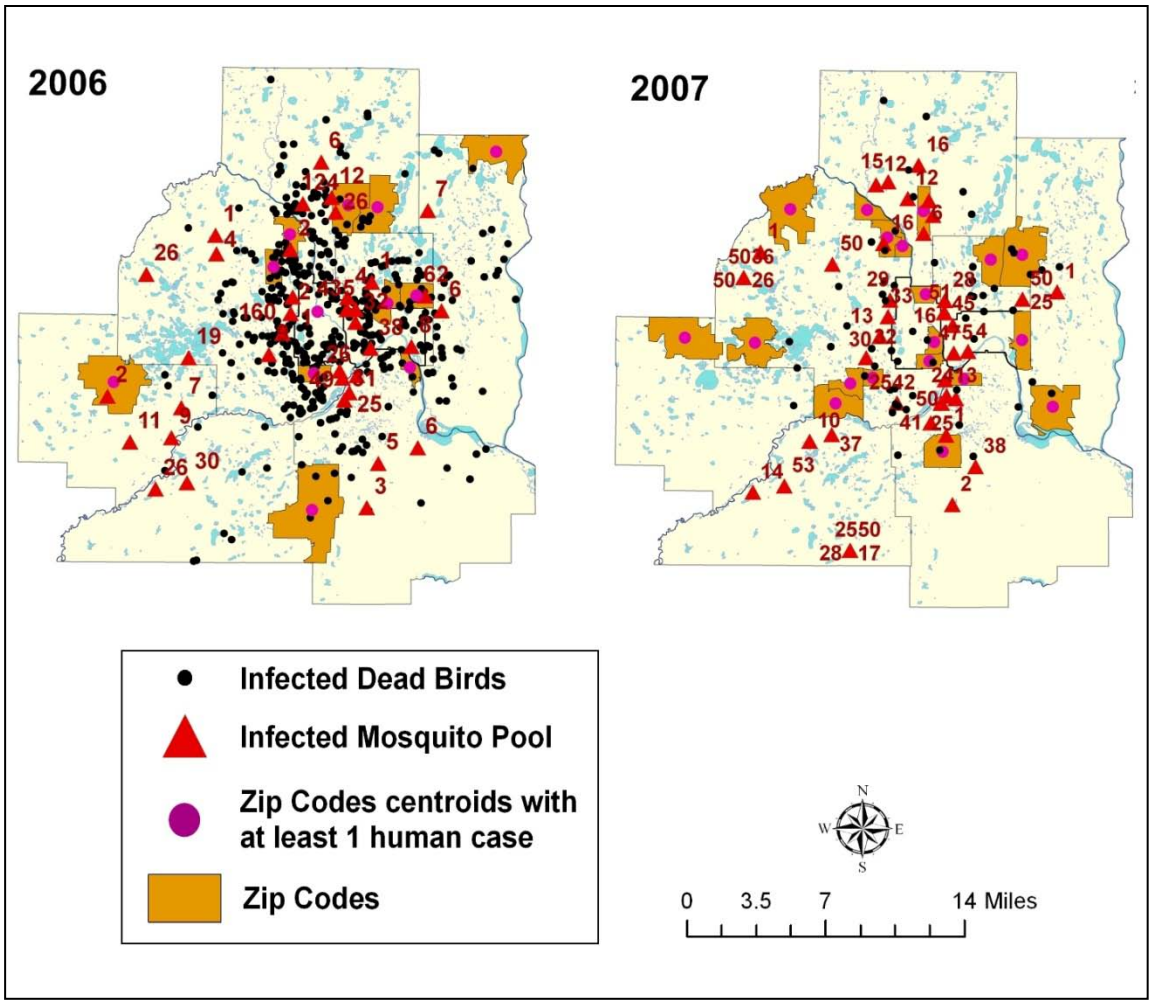


Figure 8 Sources are Minnesota Department of Health and Metropolitan Mosquito Control District. The maps show the distribution of human cases aggregated by zip codes, with their centroids as the points of occurrence, location of infected dead birds, and positive mosquito pools.

1.4 Previous Research

Since the emergence of WNV in the Western Hemisphere, especially North America, the virus has received renewed attention by researchers from various fields. The limited knowledge of how the virus propagates via complex interrelationships between human, avian, and mosquito habitat systems coupled with natural and anthropogenic risk factors prompted research on two major fronts. One such research front focused on laboratory experiments which investigated the ecological and biological potential of mosquitoes and birds to acquire and transmit the virus (Komar 2003). Studies were also conducted on different WNV strains and changing nature of virulence.

The other major research initiative focused on *geography* or spatial distribution of WNV cases. Studies mainly focused on spatio-temporal trend of the virus spread, interaction of birds, mosquitoes, and animals that could result in the transmission of the virus to humans in heterogeneous environments, and identify natural and man-made risk factors that could trigger WNV amplification. Several disciplines namely epidemiology, environmental health, and geography have contributed significantly in this research focus. The following paragraphs briefly summarize some of the important trends in WNV research. Detailed review of literature is presented in the later chapters.

There are two dominant trends in the geographic research of WNV. First, there are several attempts to delineate *exposure areas* so that public health officials could intervene with appropriate control measures and reduce the risk of infection in humans (Mostashari et al. 2003b; Theophilides et al. 2003; Eidson et al. 2005; Corrigan et al. 2006; Johnson et al. 2006; Tachiiri et al. 2006; Theophilides et al. 2006). A second trend includes modeling the WNV dynamics to understand the relationships between disease occurrence and risk factors. This would allow identifying potential *risk factor*(s) whose management would result in effective disease prevention and containment (Brownstein et al. 2002; Ruiz et al. 2004; Bowman et al. 2005; Cooke, Katarzyna et al. 2006; Diuk-Wasser et al. 2006; Gibbs et al. 2006; David et al. 2007; Lian et al. 2007).

Various approaches were used to identify WNV exposure areas, both prospectively and retrospectively. The techniques used *either* unusual sightings of infected dead birds or human cases. The existing methodologies are reviewed critically in Chapter 3. However a significant and largely unmet need is to incorporate the temporal characterization of virus spread and locational information of all the three components of transmission cycle, including birds, mosquitoes, and humans on a localized scale. Unlike the previous techniques, I believe, exposure areas containing all the three components of WNV cycle in close proximity have higher potential to amplify an outbreak as compared to exposure areas delineated by a single component only.

It is challenging to understand the spread of WNV because it propagates via complex interrelationships between human, avian, and mosquito habitat systems coupled with risk factors. The four broad categories of risk factors underlying WNV incidences are: environmental (temperature, precipitation, vegetation, hydrologic features, parks), socioeconomic (occupation, income, housing age and condition), built environment (catch basins, construction sites, ditches, scrap-tire stockpiles, sewers), and existing mosquito abatement policies. Previous models built to understand the relationships between disease occurrence and potential risk factors assumed *a priori* that there is a linear relationship between these risk factors and WNV incidences (Brownstein et al. 2002; Ruiz et al. 2004; Bowman et al. 2005; Cooke, Katarzyna et al. 2006; Diuk-Wasser et al. 2006; Gibbs et al. 2006; David et al. 2007; Lian et al. 2007). A detailed literature review of the models used to extract the relationships between the risk factors and disease occurrence is presented in Chapter 7. It is difficult for linear models to incorporate the complexities of WNV transmission network. As a result, previous investigations were not able to rigorously characterize the causal factors of the infection. It is only by simultaneously examining all of the factors as well as their interplay that we can hope to understand WNV transmission network.

From a public health issue, both these research trends have become increasingly important because the virus is spreading rapidly and causing infection among hundreds and thousands of birds and humans. Further, the recent unexpected outbreaks of WNV

in urban populated areas in the western hemisphere with *Culex* species, a predominantly house mosquito, being the important carrier of the virus marked this virus as an important emerging human pathogen. Clearly much research remains to be done on this dominant emerging infectious disease.

1.5 Goals and Research Questions

Broadly, the goal of this dissertation is to develop novel approaches that address methodological and conceptual shortcomings of our understanding of WNV and determine *why*, *when*, and *where* the virus strikes in the TCMA. The specific research questions are as follows.

1. Where are the exposure areas of WNV in the TCMA?
2. Does the spatial distribution of WNV infected dead birds, positive mosquito pools, and human cases show clustering in the urban areas of Minneapolis and Saint Paul?
3. How do urban landscape features in the TCMA associate with the transmission of the virus?
4. What are the contributing risk factors and how are they related to the occurrence of WNV incidences in the TCMA?

1.6 Outline of the Dissertation

This section briefly outlines the various topics considered in this thesis. Chapter 2 introduces the study area of TCMA. This chapter also describes the structure of WNV database, which stores georeferenced data on WNV infected dead birds, mosquito

pools, human cases, and hypothesized risk factors. The source, data type, and preprocessing techniques for each datasets are described in details. Chapter 3 then attempts to answer the first research question by developing a novel space-time interaction model to identify transmission cycles as exposure areas at various scales in the TCMA. The model, Nearest-Neighbor-Distance-Time or NNDT is a combination of geographic principles and ecological knowledge of WNV. It uses non-random spatiotemporal distribution of location of infected dead birds, mosquito pools, and WNV infected human cases to identify exposure areas.

Moving from the analysis of WNV incidences, Chapter 4 focuses on the potential risk factors. This chapter investigates the spatial association between the disease occurrences and the potential drivers or risk factors of WNV. The risk factors are grouped into four broad categories of environmental (temperature, wetlands, lakes, etc), built-environment (housing density, urban catch basin, etc), proximity (distance to open green space, distance to wetlands, distance to waste water discharge points), and existing vector control measures (larviciding and adulticiding). The *t-test* was conducted to identify risk factors with statistically different mean values in zip codes with WNV infected human cases than in zip codes without any reported case. This exploratory spatial data analysis (ESDA) of the relationships between the risk factors and disease occurrence formed the basis for building predictive models in the latter chapters.

Building upon Chapter 4, Chapter 5 addresses the second and the third research questions. Chapter 5 identifies urban landscape features of TCMA that contributed to the viral activities of WNV from the year 2002 to 2006. To accomplish this, an urban factorial ecology approach was followed to divide the TCMA into concentric urban classes. These classes were City-High Density or urban core in the center, followed by City-Medium Density, Suburb, Outer Suburb 1, and lastly Outer Suburb 2. The WNV incidence rates in birds, mosquitoes, and humans were associated with these urban classes. Finally specific urban features capable of providing habitats for birds and mosquitoes were identified.

The next two chapters concentrate on building and interpreting the WNV analysis model to understand the relationships between important risk factor and disease occurrence. The specific objective of Chapter 6 is to describe the procedures involved in building a nonlinear computational neural network model (CNN). The model captures the *nonlinear* relationships between the hypothesized risk factors and WNV incidences. Initially a Genetic Algorithm (GA) was used to search the predictor space for the best subset of predictor variables that were then included in a CNN model. The results obtained from the CNN model were also compared to that of the Ordinary Least Square (OLS) regression model for assessment of model quality. Almost all of the model quality assessments and external cross-validation results with new datasets indicated that the CNN model had better predictive capabilities than the OLS model.

Chapter 7 describes two methods that were adapted from the computational chemistry literature to interpret the neural network WNV model, using broad and detailed interpretations. Broad interpretation is a sensitivity analysis, which provides a quantitative measure of predictor (risk factor) importance in a neural network. However, this method has limited abilities, since it only provides information about which risk factor is the most important for the model's predictive ability. It does not provide any insight into the nature of the correlation between the input to the network and the output from the network. A method to extract detailed relationships regarding the risk factor and disease occurrence encoded in the weights and biases of a trained neural network models is described in Chapter 7. This method is analogous to the Partial Least Squares (PLS) interpretation technique for linear models. The method simplifies the neural network and considers the hidden neurons of the network in a manner analogous to the latent variables of the PLS interpretation. Analyzing the weights and biases, the method presented is able to provide a detailed view of the correlations between the input risk factors and the predicted value of infected dead birds. Finally, Chapter 8 summarizes the results of the studies presented in this work and concludes by highlighting the vector control policies and contributions of this thesis to the field of health geography and GIS.

2. Chapter 2: Preliminaries: Study Area and Description of Data

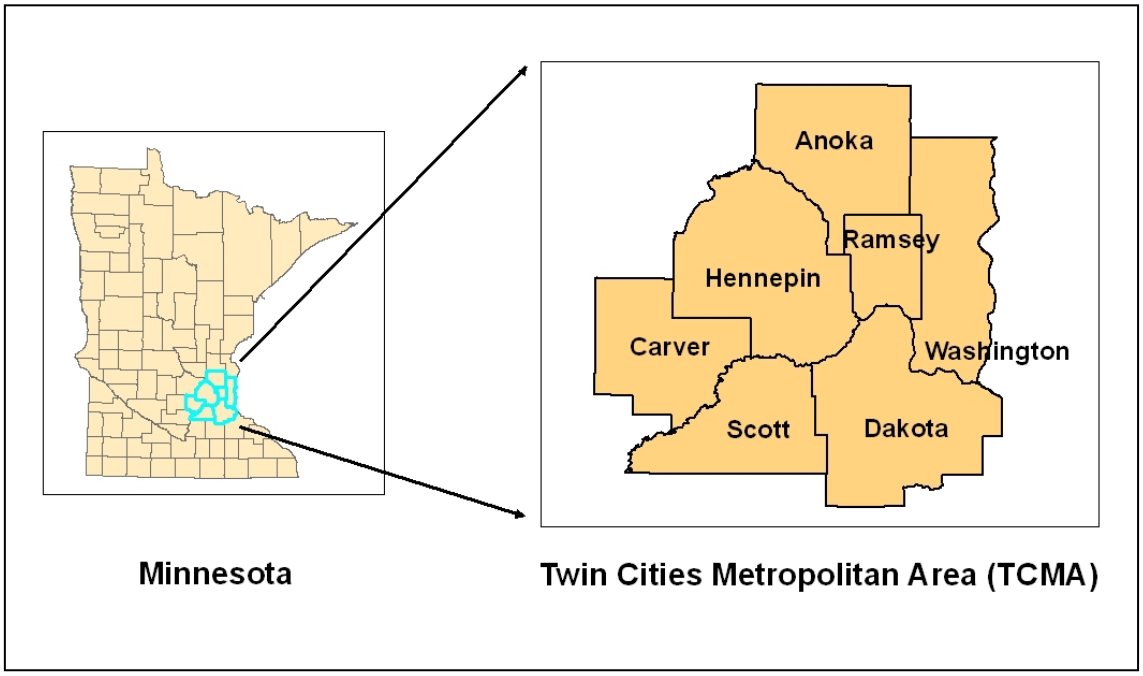
There were two important preliminary steps to be conducted. First, I selected Twin Cities Metropolitan Area of Minnesota (TCMA) as the study area to investigate the dynamics of WNV. The metropolitan area is comprised of seven counties, including Anoka, Hennepin, Carver, Scott, Dakota, Ramsey, and Washington. Second, I constructed a WNV database including infected cases and hypothesized risk factors. The four broad categories of risk factors underlying WNV incidences are: environmental (temperature, precipitation, vegetation, hydrologic features, etc.), built environment (catch basins, ditches, housing density, sewers, etc.), proximity (distance to lakes, bogs, swamps, etc.), and existing mosquito abatement programs. The section 2.1 describes the study area and section 2.2 details the structure and construction of the WNV database. The sources, data type, and preprocessing stages for incidence data and potential risk factors are described in the section 2.3 and section 2.4 respectively.

2.1 Study Area

The study area is Twin Cities Metropolitan Area of Minnesota (TCMA) including the seven counties of Anoka, Hennepin, Carver, Scott, Dakota, Ramsey, and Washington. This 7,700 km2 seven-county area is the economic hub of a multistate

region. Home to 2.8 million people, it is forecasted to top 3.5 million by 2020. It is also a major center of sprawl due to the rapid expansion of low-density suburbs into formerly rural areas and the creation of urban, suburban, and exurb agglomerations buffered from others by undeveloped land.

Figure 9 Study area of Twin Cities Metropolitan Area of Minnesota



The urban areas in the TCMA are characterized by significant presence of built area as well as patches of natural areas in the form of lakes, parks, wetlands, golf courses, and trails. These natural areas can provide habitat for birds as well as for mosquitoes. In addition, some of the features of built-environment such as catch basins, storm water ponds, construction sites, stock pile of abandoned tires, and swimming pools in the backyards of residential houses can also provide potential breeding grounds for mosquitoes and therefore increase the risk of WNV. This combination of natural and built-environment features creates an urban heterogeneous environment suitable for bird

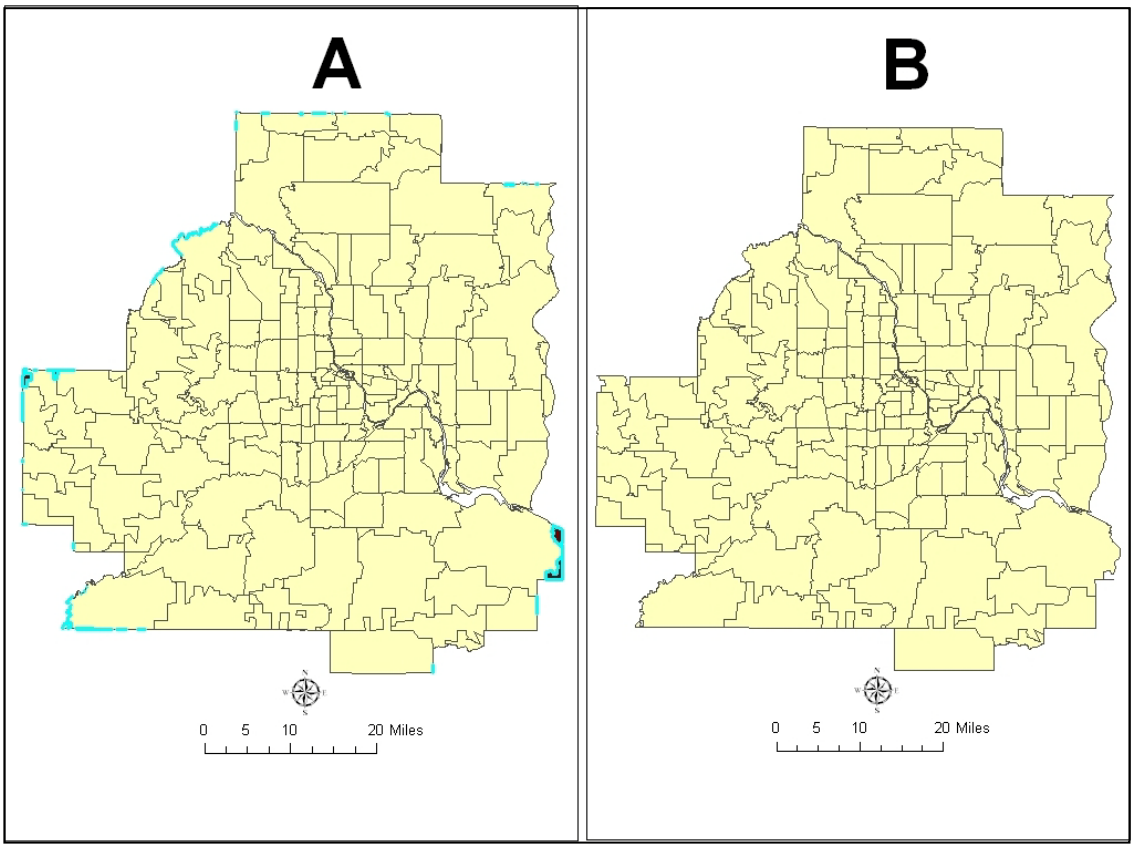
and mosquito habitats as well as their interaction. This is especially important in case of TCMA, because two out of four WNV carrying vectors, *Culex restuans* and *Culex pipiens* are predominantly urban mosquitoes.

The virus first reached Minnesota in 2002. By 2003, WNV infections resulted in several hotspots and 2007 saw one of the severest incidences of WNV in Minnesota (see section 1.3 and sub section 1.3.1). The occurrence of warmer and wetter falls in Minnesota suggests that mosquito life span may increase, leading to an increased incidence of WNV infections in the future. Further, the nature of WNV outbreak in the TCMA depicted similar trends that were observed during the three recent WNV epidemics in southern Romania, the Volga delta in southern Russia, and the northeastern United States between 1996 and 1999 (Hayes 2007). These trends are as follows. First, in all of the three epidemics mentioned above, *Culex pipiens*, the common house mosquito, was the principle vector species to transmit the virus to humans. Likewise the *Culex pipiens*, along with *Culex restuans* and *Culex tarsalis* are the main carriers of WNV in the TCMA. The researchers at Minnesota Department of Health (MDH) and Metropolitan Mosquito Control District (MMCD) indicated that *Culex restuans* appear to be important in circulating the virus between birds and mosquitoes in early summer and *Culex pipiens* then play a major role in amplifying the virus later in the season. Second, in the Midwestern United States, following Chicago and Detroit, the TCMA experienced higher incidences of WNV infection in birds, mosquitoes, and humans (see sections 1.3.1). Third, the emergence of WNV in the United States resulted in a significant decline of bird population (Koenig et al. 2007). This trend is also reflected in the TCMA. As of 2007, the number of dead bird reports due to WNV was as high as 1,111 cases (Table 3). Therefore the appearance of WNV infection in the TCMA became a major public health concern for MDH and MMCD. Both these organizations are responsible for surveillance and monitoring of WNV infected birds, vector population, equine, and human cases.

The MDH releases human WNV surveillance data *only* for research purposes aggregated at the US Census zip code level (see section 2.3 for details). Therefore the

smallest unit of analysis of this study is zip codes. Of the total 864 zip codes in Minnesota, 176 are present in the TCMA. Further, the zip codes in the TCMA were checked for slivers and discontinuous polygons (zip codes). The final study area covering the seven counties metropolitan area is comprised of 159 zip codes. Figure 10 shows the study area with zip code boundaries. Part A shows all of the 176 zip codes boundaries, of which the zip codes with discontinuous boundaries and slivers are highlighted. These highlighted polygons were deleted from the final study area, shown in Part B.

Figure 10 The boundaries of US Census zip codes of the Twin Cities Metropolitan Area of Minnesota



2.2 West Nile virus database

The WNV database stored georeferenced data of infected birds, mosquitoes, and human cases. Data on hypothesized risk factors were also included in the database. On the basis of thorough literature review of previous studies on WNV and my domain knowledge of the research problem, the factors underlying the occurrence of WNV incidences were broadly divided into four categories. The categories are: environmental (temperature, precipitation, vegetation, hydrologic features, etc.), built environment (catch basins, ditches, housing density, sewers, etc.), proximity (distance to lakes, bogs, swamps, etc.), and existing mosquito abatement programs. The inclusion of specific risk factors was also consulted with an epidemiologist, David Neitzel from the MDH[1] and vector ecologist, Kirk Johnson from the MMCD[2].

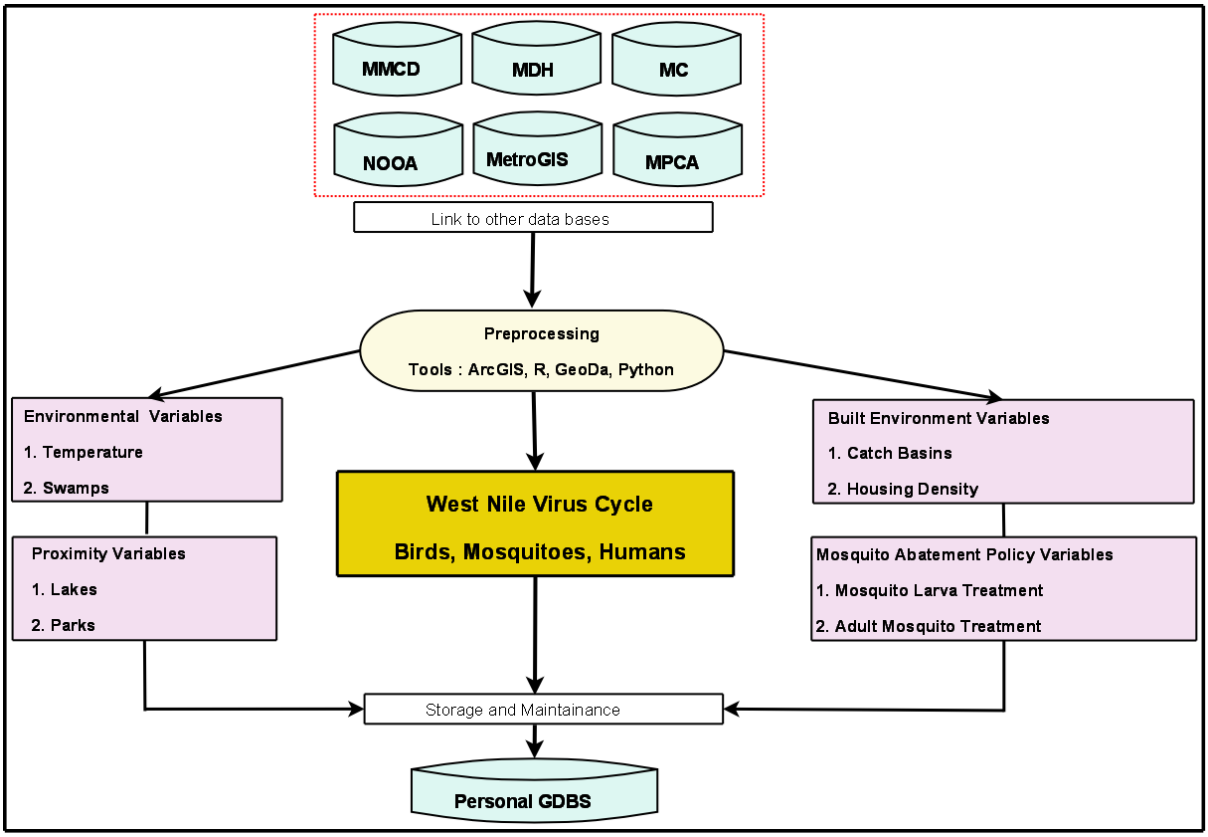
Georeferenced data of all the factors and WNV infected cases were obtained from secondary sources such as MDH, MMCD, Minnesota Department of Natural Resources (DNR), and MetroGIS. The data formats were typically in ESRI shape file and grid formats (ESRI). Several tools and software packages, such as, ArcGIS 9.x (ESRI), Hawth's Tools (Hawthorne), X-Pro tools (X-Pro), GeoDa (GeoDa), R statistical software (R), and Python programming language were used for processing the data into common projection system and spatial resolution. The Universal Transverse Mercator (UTM) projection with Zone 15 was used as the common projection system. As stated previously the smallest unit of analysis or spatial resolution was US Census zip codes. Finally, all the data were stored in a personal geodatabase, named as 'West Nile virus database of the Twin Cities Metropolitan Area of Minnesota (WNV_TCMA)'. The database was maintained and updated regularly as new data were obtained. The structure of the database is shown in Figure 11.

---

[1] David Neitzel is an epidemiologist at the Minnesota Department of Health. He works in the departments of Infectious Disease Epidemiology and Acute Disease Investigation and Control. Freeman Office Building, 625 N. Robert Street, P.O. Box 64975. Saint Paul, MN 55164 – 0975. Ph: 651-2015414. Email: David.Neitzel@state.mn.us

[2] Kirk, A. Johnson, vector ecologist at the Metropolitan Mosquito Control District. 2099 University Avenue West, Saint Paul, MN 55104. Ph: 651-6438370. Email: kjohnson@mmcd.org

Figure 11 Design of West Nile Virus Database



2.3 WNV Incidence Data

For WNV incidence data, the study used three datasets for the years 2002 to 2007. These datasets are as follows:

1. **Infected dead birds**

   The MMCD provided location and date of WNV infected dead birds (American Crow and Blue Jay), which were typically reported by people. The raw data were later geocoded at the street level, verified, and cleaned with occasional field validation by MMCD staff.

2. **Infected Mosquito Pools**

   The location of WNV positive mosquito pools were also provided by the MMCD. These were mosquito traps designed to collect mosquitoes of different species for viral analysis on a weekly basis (MMCD 2004). MMCD has three types of collection traps for mosquitoes, which are distributed throughout the metropolitan area. The trap types are CO2 traps, gravid traps, and sweep nets. The CO2 traps are elevated in the tree canopy and are used for collecting female mosquito samples in their host-seeking phase. The gravid traps located on the ground are designed to attract female mosquitoes that are seeking oviposition sites (i.e., places to lay eggs). Mosquito pool data included attributes such as collection date, specie type, and number of infected mosquitoes in the pool.

3. **Infected Human cases**

   The MDH provided the location and the onset date of infected human cases by zip code. Due to confidentially of human health-related data and the guidelines of the University of Minnesota's Institutional Review Board (IRB), WNV-infected human cases were obtained aggregated at the zip code. The IRB determined that this study (reference number 0611E96307) is exempted from the

review under federal guidelines 45 CFR Part 46.101(b) category #4 including existing data, records review, and pathological specimens.

Table 4 summarizes the WNV incidence data.

Table 4 Sources and units of West Nile Virus Incidence Data

Data	Details	Source
Infected Dead Birds	Point data, reported date	MMCD
Mosquito Pools	Point data, reported date	MMCD
Human cases	Areal data (zip codes), reported date	MDH

2.4 WNV Risk Factor Data

As mentioned above, WNV risk factors are broadly divided into four categories: environmental, built-environment, proximity, and existing vector control programs. The following sections describe the source, data type, and preprocessing stages of all the risk factor variables considered in this study. The variables that had time series data were processed for the years 2002 to 2007.

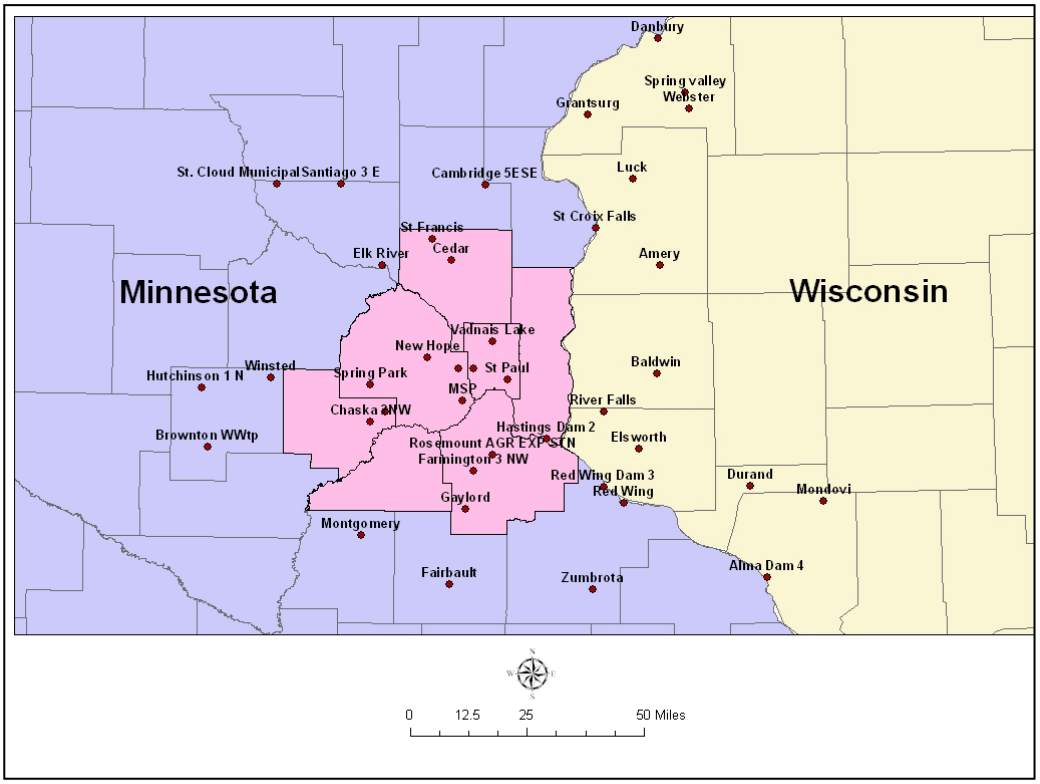
2.4.1 Environmental Factors

This section includes the description of all the environmental factors associated with the occurrence of WNV incidences in birds, mosquitoes, and humans. The risk factors are as follows:

1. **Average Maximum Daily Temperature (Avg_Max_Temp)**

Maximum daily temperature data was obtained from the Cooperative data of Climatological Observation Stations (COOP). This data is published by the National Climatic Data Center (NCDC), NOAA Satellite and Information Service, and Minnesota Climatology Working Group. The location of observation stations considered in this study are shown in Figure 12.

Figure 12 Weather Observation Stations in the Twin Cities Metropolitan Area and surrounding region in Minnesota and Wisconsin



The average of seven days daily maximum temperature prior to the reporting dates of WNV-infected dead birds in 2006 was calculated as the final Avg_Max_Temp variable. The details of data processing are as follows.

a. Daily maximum temperature data of seven days before the reporting date of an infected dead bird was obtained from the weather station nearest to the location of that particular dead bird.
b. 7-day average of maximum daily temperature was calculated
c. This was repeated for all the infected dead birds reported in the year 2006.
d. The final variable was the average of 7-day average of daily maximum temperature associated with all the birds reported in a zip code.
e. The average of 7-day average of daily maximum temperature was then interpolated for all the other zip codes with no reports of dead birds.

2. Average Minimum Daily Temperature (Avg_Min_Temp)

The data sources and methodology to create the average daily minimum temperature variable was same as that of the average daily maximum temperature.

3. Average Daily Precipitation (Avg_Precip)

The data sources and procedure to create the average daily precipitation variable was same as that of average daily maximum temperature.

4. Percentage area of Lakes (Perct_Lakes)

The Public Waters Inventory (PWI) distributed by the Minnesota Department of Natural Resources (MN DNR) was used as the data for lakes. Here the term lake is generic, describing any water body greater than 3 acres. The final variable, percentage area of lakes (Perct_lakes) for each zip code, was calculated by the following formula. The ‘Polygon in Polygon Analysis’ tool

was used from the Hawths' Analysis Tools for ArcGIS (Hawthorne).

$$Perct\_Lake_i = \frac{\sum area\ of\ lakes\ \ in\ zip\ code\ (i)}{Area\ of\ zip\ codes\ (i)} * 100$$

5. Percentage of Wetland Area (Perct_area_WT)

The data on wetlands was acquired from the classification of mosquito breeding sites compiled by the MMCD. The MMCD breeding site system classifies sites according to habitats and is based on the classification systems of the U.S. Army Corps of Engineers (1978) and the United States Fish and Wildlife Service (USFWS), Circular (1956, 1971). Following are the descriptions of the eight major wetland types that are found in the TCMA.

a. Type 1 – Seasonally Flooded Basins and Flat

- Vegetation: Herbaceous (non-woody and dying at the end of a season) plants and upland grasses. No canary grass. No cattails.
- Site Status: Temporary water. Usually well drained during much of the growing season. Uncultivated farmland and woodland pools are typical examples.
- Water Depth: 6 inches +/- when wet

b. Type 2 – Inland Fresh Meadow

- Vegetation: Canary grass (3 inches +/- high) and sedges.
- Site Status: Temporary water. The soil is usually without standing water during most of the growing season but is waterlogged within at least a few inches of its surface. Site is normally dry in late summe**r**
- Water Depth: 6 inches to 18 inches when wet

c. Type 3 – Inland Shallow Fresh Marsh

- Vegetation: Succession of cattails, sedges, and canary grass (at the perimeter) make up the predominant vegetation.
- Site Status: Temporary water. May contain water up to midsummer, dry up completely, or remain waterlogged for the entire season.
- Water Depth: 6 to 24 inches

d. Type 4 – Inland Deep Fresh Marsh

- Vegetation: Vegetation is in a "band" configuration surrounding or adjacent to a permanent open body(s) of water located at some areas(s) in the site. Vegetation bands include: canary grasses at the perimeter, sedges in the shallow water, followed by cattails and broadleaf plants bordering the open water. A partial vegetative mat may be present.
- Site Status: Permanent water habitat. Water present year-around. Pockets of open water allow submerged aquatic plants to grow.
- Water depth: 6 inches to 3 feet
- A subtype with permanent water status and steeper banks is not associated with mosquito development. They are marked as non-breeding ponds.

e. Type 5 – Inland Fresh Open Water

- Vegetation: Vegetation is usually cattails and broadleaf plants surrounding or adjacent to deeper open water with canary grasses and sedges being found at times in the shallower water near the perimeter.
- Site Status: Permanent water habitat containing game fish. This would include any site that is directly connected to a lake, creek, stream, or river that is designated on township plat maps.

- Water Depth: +/- 10 feet deep fringed by a border of emergent vegetation.

f. Type 6 – Shrub Swamp

- Vegetation: Includes alders, willows and dogwood as well as some herbaceous plants
- Site Status: Soil is usually waterlogged during the growing season. Sites can be isolated or occur along sluggish streams and flood plains.
- Water Depth: Usually shallow.

g. Type 7 – Wooded Swamp

- Vegetation: Northern conifer swamps can contain tamarack, white cedar, black spruce, and balsam fir.
- Site Status: The soil is waterlogged to within at least a few inches of its surface during the growing season and is sometimes covered with water
- Water Depth: Varies
- Generally this site type of wetland is considered uncommon or an exception and/or specific to certain areas of the TCMA

h. Type 8 – Bog

- Vegetation: Moss, sedges, cotton grass, and heath shrubs. Stunted black spruce and tamarack are often found in bogs.
- Site Status: The soil is waterlogged, poorly drained, and supports a spongy covering of moss.
- Water Depth: Varies.

The wetland data were further filtered based on two criteria 1) wetland area greater than one acre and 2) wetlands that were treated for *Culex* species mosquitoes in the year 2006. Finally the percentage area of each wetland type was calculated by the following formula and using the 'Polygon in Polygon

Analysis' tool from the Hawths' Analysis Tools for ArcGIS (Hawthorne).

$$Perct\_Area\_WT_{ij} = \frac{\sum area\ of\ Type_j\ \ in\ zip\ code\ (i)}{Area\ of\ zip\ codes\ (i)} * 100$$

where, j ranges from 1 to 8 representing the different wetland types in the TCMA.

6. Percentage area of Open Green Space (Perct_OpenSpace)

This variable was compiled from The Lawrence Group (TLG) landmark feature data (version March 2008). The landmark feature data are a combination of a variety of geographic and reference point features from three datasets. These datasets are geographic areas of interest (tlglmk_areas), linear features (tlglmk_lines), and points of interest (tlglmk_points). Typical features of the tlglmk_areas layer are parks, lakes, golf courses, cemeteries, airports, etc. Typical features of the tlglmk_lines layer include railroads and creeks and typical features of the tlglmk_points layer are city halls, schools, hospitals, etc. The datasets are provided by TLG and distributed by MetroGIS in an Arc/Info export format.

The open green space variable was derived from the combination of three landmark areas including national parks (CODE: D83), state/local parks (CODE: D85), and arboretum/nature centers (CODE: 87). The final variable percentage area of Open Green Space (Perct_OpenSpace) was calculated by the following algorithm. The 'Polygon in Polygon Analysis' tool was used from the Hawths' Analysis Tools for ArcGIS (Hawthorne).

$$Perct\_OpenSpace_i = \frac{\sum area\ of\ open\ green\ spaces\ in\ zip\ code(i)}{Area\ of\ zip\ codes\ (i)} * 100$$

7. Percentage area of 14 land cover variables (Perct_LC)

The 2001 National Land Cover Data (NLCD 2001) was used to extract land cover information for each zip code. The land cover data was in a raster format and distinguishes 16 different classes, of which 14 were present in the TCMA. The 14 land cover classes with their codes are as follows: Open water (11), Developed open space (21), Developed low intensity (22), Developed medium intensity (23), Developed high intensity (24), Barren land (31), Deciduous forest (41), Evergreen forest (42), Mixed forest (43), Scrub/Shrub (52), Pasture/Hay (81), Cultivated crops (82), Woody wetland (90), and Emergent herbaceous wetland (95).

To calculate these 14 land cover variables for each zip code, the following steps were followed:

a. The 'Thematic Raster Summary Tool' of Hawth's Analysis Tools for ArcGIS (Hawthorne) was used to summarize the number of cells of each category by zip code. The output was a number of cells for each category of land cover in each polygon (zip code).
b. To convert this to area, the data was transferred to R statistical programming language (R). The number of cells was multiplied by the area of cell of the original land cover raster. For example, if the cell size was 30m, and the count was 5 cells for a particular category, then the area was: (30*30)*5 = 4500m^2 (4500 square meters).
c. The area unit was then converted to square miles for compatibility with other variables in the study.
d. Finally percentage of area of 14 land cover variables were calculated by the following algorithm, where *j* ranges 1 to 14 land cover classes

$$Perct\_LC_{ij} = \frac{Area\ of\ LC_j\ in\ zip\ code\ (i)}{Area\ of\ zip\ code\ (i)} * 100$$

8. Average Elevation (Avg_Elev)

To calculate the average elevation of zip codes, I have used the raster data of Hydrologically Corrected 10 Meter Digital Elevation Model. The data were downloaded from the MetroGIS. "Hydrologically Corrected" means that the DEM is adjusted to accurately depict surface water drainage, even when the drainage is in underground storm sewers. The steps to create the average elevation (Avg_Elev) are as follows:

a. 'Zonal Statistics as Table' from the spatial analyst tool box in ArcGIS 9.x was used to calculate the average elevation for each zip code. Here the zip codes were specified as 'zones'.
b. The table output from the previous step was joined to the zip code layer and the field indicating average elevation or DEM value was retained as the final field.

9. Density of streams (Den_stream)

The stream network data was obtained from the MetroGIS. This georeferenced data was created by the collaborative efforts of MetroGIS, MN DNR, and Minnesota Department of Transportation (MnDot). The data depicts streams captured from USGS 1:24,000, 7.5 minute quadrangle maps, with perennial vs. intermittent classification, and connectivity through lakes, rivers, and small wetland basins. Streams are depicted as single lines. Attributes describe stream type, source of data, and connectivity. The final field of Den_stream per square mile was calculated by the following formula and using the 'Sum Length of Lines in Polygons' tool from the Hawths' Analysis Tools for ArcGIS (Hawthorne).

$$Den\_streams_i = \frac{\sum Length\ of\ streams\ in\ zip\ code\ (i)}{Area\ of\ zip\ code\ (i)}$$

10. Average Dew Point (Avg_dew)

Arc/Info ASCII grids at a resolution of 4 km^2 representing monthly dew point temperatures were acquired from the Oregon State University PRISM website (PRISM). The PRISM (Parameter-elevation Regressions on Independent Slopes Model) climate mapping system is a knowledge-based system that uses point measures of climatic data, digital elevation data, and expert knowledge of complex climatic extremes to produce continuous climatic surfaces for the United States. Along with the average dew point, the PRISM group also provides data on precipitation, average maximum temperature, average minimum temperature, standardized precipitation index, and percent of normal precipitation. In the year 2006, the reporting of WNV infected dead birds started from the month of June and ended in the month of September. Based on this temporal window, I downloaded ASCII grids of monthly dew point for the months of June, July, August, and September. The grids were then processed to the final variable of Avg_dew by the following procedure.

a. Downloaded grids (June, July, August, September) of dew point as .gz files
b. Extracted the compressed files and saved as text files
c. Using ArcGIS 9.x, ASCII files were converted to rasters
d. Defined projection and extracted the rasters to the study area of TCMA
e. Averaged all the four rasters to create a final raster representing the average dew point temperatures for the 2006 WNV season.
f. 'Zonal Statistics as Table' from the spatial analyst tool box in ArcGIS 9.2 was used to calculate the average dew point for each zip code. Here the zip codes were specified as 'zones'.
g. The table output from the previous step was joined to the zip code layer and the field indicating average dew point value was retained as the final field.

Table 5 summarizes the environmental risk factors of WNV incidence.

Table 5 Sources and units of environmental risk factors of West Nile virus infection

Variables	Unit	Source
Avg_Max_Temp	F	NOAA, UMN Climate Group
Avg_Min_Temp	F	NOAA, UMN Climate Group
Avg_Precip	Inches	NOAA, UMN Climate Group
Perct_Area_WT(1)	%	MMCD
Perct_Area_WT(2)	%	MMCD
Perct_Area_WT(3)	%	MMCD
Perct_Area_WT(4)	%	MMCD
Perct_Area_WT(5)	%	MMCD
Perct_Area_WT(6)	%	MMCD
Perct_Area_WT(7)	%	MMCD
Perct_Area_WT(8)	%	MMCD
Perct_OpenSpace	%	TLG, MetroGIS
Perct_lakes	%	MN DNR
Perct_LC(11)	%	NLDC 2001
Perct_LC(21)	%	NLDC 2001
Perct_LC(22)	%	NLDC 2001
Perct_LC(23)	%	NLDC 2001
Perct_LC(24)	%	NLDC 2001
Perct_LC(31)	%	NLDC 2001
Perct_LC(41)	%	NLDC 2001
Perct_LC(42)	%	NLDC 2001
Perct_LC(43)	%	NLDC 2001
Perct_LC(52)	%	NLDC 2001
Perct_LC(81)	%	NLDC 2001
Perct_LC(82)	%	NLDC 2001
Perct_LC(90)	%	NLDC 2001
Perct_LC(95)	%	NLDC 2001
Avg_Elev	Meter	MetroGIS
Den_Stream	per/sq mile	MetroGIS
Avg_dew	F	PRISM

2.4.2 Built-Environment Factors

The built-environment factors, which are often created by human activities, also increase the risk of WNV infection. The built-environment factors included in this study are as follows.

1. **Catch Basins and Storm Water Ponds**

   The data on spatial location of urban catch basins and storm water ponds was acquired from the MMCD. They mapped the location of catch basins in the entire seven county regions in 2006. Two variables were derived from this data: density of all catch basins per/sq mile (Den_cb) and density of catch basins with stagnating water per/sq mile (Den_wcb)

   a. **Den_cb**

   $$Den\_cb_i = \frac{Count\ of\ catch\ basins\ in\ zip\ code\ (i)}{Area\ of\ zip\ code\ (i)}$$

   b. **Den_wcb**

   $$Den\_wcb_i = \frac{Count\ of\ catch\ basins\ with\ water\ in\ zip\ codes(i)}{Area\ of\ zip\ code\ (i)}$$

2. **Density of Ditches (Den_ditch)**

   Ditches data were obtained from the MnDot. This layer consists of a number of individual data layers or themes digitized from USGS 7.5-minute quadrangles. They fall into the following broad categories of transportation system, civil and political boundaries, and surface water. The final variable, density of ditches (Den_ditch) per/sq mile was created by the following algorithm and using the 'Sum Length of Lines in Polygons' tool from the Hawths' Analysis Tools for ArcGIS (Hawthorne).

$$Den\_ditch_i = \frac{\sum Length\ of\ ditches\ in\ zip\ code\ (i)}{Area\ of\ zip\ code\ (i)}$$

3. Percentage Area of Impaired Lakes (Perct_Imp_lakes)

The federal Clean Water Act (CWA) requires states to adopt water-quality standards to protect lake waters from pollution. These standards define how much of a pollutant can be in the water and still allow it to meet designated uses, such as drinking water, fishing, and swimming. The standards are set on a wide range of pollutants, including bacteria, nutrients, turbidity and mercury. A water body is "impaired" if it fails to meet one or more of the water quality standards. Minnesota Pollution Control Agency (MPCA) is responsible for assessing the lakes and listing impaired ones. The agency also coordinates closely with other state and local agencies on restoration activities. The lakes determined as 'impaired' by MPCA in 2006 were used to create this variable. The percentage of area of impaired lakes (Perct_Imp_lakes) is calculated by the following formula. The 'Polygon in Polygon Analysis' tool was used from the Hawths' Analysis Tools for ArcGIS (Hawthorne).

$$Perct\_Imp\_lake_i = \frac{\sum area\ of\ imapired\ lakes\ in\ zip\ code(i)}{Area\ of\ zip\ codes\ (i)} * 100$$

4. Density of sewers (Den_sewers)

Owned, operated and maintained by the MetroGIS, the sewer interceptor system provides the link from community sewer systems to the wastewater treatment facilities in the seven county metropolitan area. The final variable, density of sewers (Den_sewers) per/sq mile was created by the following algorithm and using the 'Sum Length of Lines in Polygons' tool from the Hawths' Analysis Tools for ArcGIS (Hawthorne).

$$Den\_sewers_i = \frac{\sum Length\ of\ sewers\ in\ zip\ code\ (i)}{Area\ of\ zip\ code\ (i)}$$

5. Housing Density (H_density)

The 2006 regional parcel dataset of the TCMA was used to calculate the housing density for each zip code. This dataset is a compilation of tax parcel polygons and point layers from the seven county TCMA. The polygon layer contains one record for each real estate/tax parcel polygon. The data are distributed by Metropolitan Council (MC) on behalf of MetroGIS.

In the first step, parcels (both residential and commercial) with estimated building value (EMV_BLDG) greater than zero were selected for further analysis. This was done to retain only valid buildings. Finally the housing density (H_density) per acre was calculated by the following formula and using the 'Polygon in Polygon Analysis' tool from the Hawths' Analysis Tools for ArcGIS (Hawthorne).

$$H\_density_i = \frac{Count\ of\ parcels\ in\ zip\ code\ (i)}{Area\ of\ zip\ codes\ (i)}$$

6. Average Housing Age (Avg_H_Age)

Similar to the housing density, average housing age (Avg_H_Age), was also derived from the 2006 regional parcel dataset of the TCMA.

7. Density of Roads (Den_roads)

The road data was acquired from TLG's street centerline and address range dataset for the Twin Cities Metropolitan Area and beyond. This is a

MetroGIS Regionally Endorsed dataset. The density of roads (Den_roads) per/sq mile was created by the following algorithm. The 'Sum Length of Lines in Polygons' tool was used from the Hawths' Analysis Tools for ArcGIS (Hawthorne).

$$Den\_roads_i = \frac{\sum Length\ of\ roads\ in\ zip\ code\ (i)}{Area\ of\ zip\ code\ (i)}$$

8. Density of Bike paths (Den_bikeways)

The regional bikeways dataset was created by the MnDOT in 2003. It is maintained and updated by the Land Management Information Center (LMIC) through a contract with the MC. Dataset includes bicycle routes within nine counties: Anoka, Carver, Chisago, Dakota, Hennepin, Ramsey, Scott, Washington, and Wright. The bikeways are from a number of sources including the Metro Bicycle Network map book (2001), supplemented by information from maps published by city, state, county and regional government agencies, and city and county planning maps.

In the first step, only existing bikeways were extracted for further consideration. The final variable, density of bikeways (Den_bikeways) per/sq mile was created by the following formula and using the 'Sum Length of Lines in Polygons' tool from the Hawths' Analysis Tools for ArcGIS (Hawthorne).

$$Den\_bikeways = \frac{\sum Length\ of\ bikeways\ in\ zip\ code\ (i)}{Area\ of\ zip\ code\ (i)}$$

9. Population Density (Pop_density)

Population density was calculated for each zip code from the US census zip code file

Table 6 summarizes the built environment risk factors of WNV incidence.

Table 6 Sources and units of built-environment risk factors of West Nile virus infection

Variables	Unit	Source
Den_cb	per/sq mile	MMCD
Den_wcb	per/sq mile	MMCD
Den_ditch	per/sq mile	MnDOT
Perct_Imp_lakes	%	MPCA
Den_sewers	per/sq mile	MetroGIS
H_density	per/acre	MC, MetroGIS
Avg_H_Age	years	MC, MetroGIS
Den_roads	per/sq mile	TLG, MetroGIS
Den_bikeways	per/sq mile	MnDOT, LMIC, MetroGIS
Pop_density	per/acre	US Census

2.4.3 Proximity Factors

This category includes average proximity to a WNV hypothesized risk factor from a zip code. Here the variables were quantified as the average Euclidean distance (in miles) to a given risk factor such as, lakes, bogs, and sparks from a zip code in the TCMA. The variables included in this category are as follows:

1. **Average distance to 8 types of wetlands (Avg_d_Type $_j$)**

The characteristics of 8 different wetlands types are described in details in section 2.4.1. The proximity to the wetlands is calculated by the following steps:

a. Using the 'Euclidean Distance' function from the spatial analyst extension of ArcGIS 9.x, a Euclidean raster to a given wetland type was created. This raster represented, for each cell of size 10 meter, the Euclidean distance to the nearest source of wetland type $_j$.

b. 'Zonal Statistics as Table' from the spatial analyst tool box in ArcGIS 9.x was used to calculate the average distance to wetlands of type $_j$ for each zip code. Here the zip codes were specified as 'zones'.

c. The table output from the previous step was joined to the zip code shape file and the field indicating average distance was retained as the final field, Avg_d_Type $_j$

d. The unit of distance was converted to miles for consistency with the other variables in this study

e. Following the steps from a to d, proximity variables were created for 8 types of wetlands, **Avg_d_Type1**, **Avg_d_Type2, Avg_d_Type3, Avg_d_Type4, Avg_d_Type5, Avg_d_Type6, Avg_d_Type7,** and **Avg_d_Type8.**

2. **Average distance to WNV infected mosquito pools (Avg_d_pools)**

The location of WNV infected mosquito pools in the year 2006 were used here (for details see section 2.3). The variable, Avg_d_pool was calculated by the following methodology:

a. Using the 'Euclidean Distance' function from the spatial analyst extension of ArcGIS 9.x, a Euclidean raster was created. This raster calculated, for each cell of size 10 meter, the Euclidean distance to the closest source of infected mosquito pool.

b. ‘Zonal Statistics as Table’ from the spatial analyst tool box in ArcGIS 9.x was used to calculate the average distance to positive mosquito pool for each zip code. Here the zip codes were specified as ‘zones’.
c. The table output from the previous step was joined to the zip code shape file and the field indicating average distance was retained as the final field, Avg_d_pools
d. The unit of distance was converted to miles for consistency.

3. Average distance to lakes (Avg_d_lakes)

The proximity to PWI lakes (Avg_d_lakes) (for details see section 2.4.1) in miles was calculated by the following steps:

a. Using the ‘Euclidean Distance’ function from the spatial analyst extension of ArcGIS 9.x, a Euclidean raster was created. This raster measured, for each cell of size 10 meter, the Euclidean distance to the closest source of lakes.
b. ‘Zonal Statistics as Table’ from the spatial analyst tool box in ArcGIS 9.x was used to calculate the average distance to lakes for each zip code. Here the zip codes were specified as ‘zones’.
c. The table output from the previous step was joined to the zip code shape file and the field indicating average distance was retained as the final field, Avg_d_lakes
d. The unit of distance was converted to miles for consistency.

4. Average distance to open green spaces (Avg_d_openspace)

The distance to open green spaces (Avg_d_openspace) (for details on ‘open green space’ see section 2.4.1) in miles was calculated by the following steps:

a. Using the ‘Euclidean Distance’ function from the spatial analyst extension of ArcGIS 9.x, a Euclidean raster was created. This raster

calculated, for each cell of size 10 meter, the Euclidean distance to the closest source of open green space.

b. 'Zonal Statistics as Table' from the spatial analyst tool box in ArcGIS 9.2 was used to calculate the average distance to open green space for each zip code. Here the zip codes were specified as 'zones'.
c. The table output from the previous step was joined to the zip code shape file and the field indicating average distance was retained as the final field, Avg_d_openspace
d. The unit of distance was converted to miles for consistency with the other variables included in this study.

5. Average distance to sewers (Avg_d_sewers)

The proximity to sewers (Avg_d_sewers) (for details see section 2.4.2) in miles was calculated by the following steps:

a. Using the 'Euclidean Distance' function from the spatial analyst extension of ArcGIS 9.x, a Euclidean raster was created. This raster represented, for each cell of size 10 meter, the Euclidean distance to the closest source of sewers.
b. 'Zonal Statistics as Table' from the spatial analyst tool box in ArcGIS 9.x was used to calculate the average distance to sewers for each zip code. Here the zip codes were specified as 'zones'.
c. The table output from the previous step was joined to the zip code shape file and the field indicating average distance was retained as the final field, Avg_d_sewers
d. The unit of distance was converted to miles for consistency with the units of other variables.

6. Average distance to waste water discharge points (Avg_d_waste)

The data on the location of waste water discharge points was obtained from the Industrial Waste and Pollution Prevention Section of the Metropolitan Council Environmental Services Division (MCES). MCES is delegated as the control authority to regulate the wastewater hauled into the service area. Using this data, the average distance to the discharge points (Avg_d_waste) was calculated by the following methodology:

a. Using the 'Euclidean Distance' function from the spatial analyst extension of ArcGIS 9.x, a Euclidean raster was created. This raster calculated, for each cell size of 10 meter, the Euclidean distance to the nearest source of waste water discharge point.
b. 'Zonal Statistics as Table' from the spatial analyst tool box in ArcGIS 9.x was used to calculate the average distance to discharge points for each zip code. Here the zip codes were specified as 'zones'.
c. The table output from the previous step was joined to the zip code shape file and the field indicating average distance was retained as the final field, Avg_d_waste
d. The unit of distance was converted to miles for consistency.

7. Average distance to streams (Avg_d_streams)

The distance to nearest stream (Avg_d_streams) (for the details see section 2.4.1) in miles was calculated by the following steps:

a. Using the 'Euclidean Distance' function from the spatial analyst extension of ArcGIS 9.x, a Euclidean raster was created. This raster calculated, for each cell size of 10 meter, the Euclidean distance to the closest source of stream.
b. 'Zonal Statistics as Table' from the spatial analyst tool box in ArcGIS 9.x was used to calculate the average distance to streams for each zip code. Here the zip codes were specified as 'zones'.

c. The table output from the previous step was joined to the zip code shape file and the field indicating average distance was retained as the final field, Avg_d_streams
d. The unit of distance was converted to miles for consistency with the units of other variables included in this study.

8. Average Distance to Golf Courses (Avg_d_golf)

The data on the location of golf courses in the TCMA was obtained from the TLG landmark data. This is a compilation of a variety of geographic and reference point features from three datasets, Geographic areas of interest (tlglmk_areas), linear features (tlglmk_lines), and points of interest (tlglmk_points). The golf courses are included in the tlglmk_areas file (CODE: D81). The average distance to the golf courses (Avg_d_golf) in miles was derived from the TLG landmark data by the following methodology:

a. Using the 'Euclidean Distance' function from the spatial analyst extension of ArcGIS 9.x, a Euclidean raster was created. This raster calculated, for each cell size of 10 meter, the Euclidean distance to the nearest source of a golf course.
b. 'Zonal Statistics as Table' from the spatial analyst tool box in ArcGIS 9.x was used to calculate the average distance to golf courses for each zip code. Here the zip codes were specified as 'zones'.
c. The table output from the previous step was joined to the zip code shape file and the field indicating average distance was retained as the final field, Avg_d_golf
d. The unit of distance was converted to miles for consistency with the other variables.

9. Average Distance to Trails (Avg_d_trails)

The trail data is acquired from the 'Regional and State Trails' layer distributed by MC. It is a compilation of alignments for existing and proposed regional and state trails for the 7 county metropolitan area. The final variable, average distance to trails (Avg_trails) in miles was calculated by the following steps:

a. Only existing trails alignments were extracted and used for further analysis
b. Using the 'Euclidean Distance' function from the spatial analyst extension of ArcGIS 9.x, a Euclidean raster was created. This raster calculated, for each cell size of 10 meter, the Euclidean distance to the closest source of trails (existing).
c. 'Zonal Statistics as Table' from the spatial analyst tool box in ArcGIS 9.x was used to calculate the average distance to trails for each zip code. Here the zip codes were specified as 'zones'.
d. The table output from the previous step was joined to the zip code shape file and the field indicating average distance was retained as the final field, Avg_d_trails.
e. The unit of distance was converted to miles for consistency.

10. Average Distance to Bikeways (Avg_d_bike)

The distance to nearest bikeways (Avg_d_bike) (for the details on 'bikeways' see section 2.4.2) in miles was calculated by the following steps:

a. Using the 'Euclidean Distance' function from the spatial analyst extension of ArcGIS 9.x, a Euclidean raster was created. This raster calculated, for each cell size of 10 meter, the Euclidean distance to the closest source of bikeways.

b. ‘Zonal Statistics as Table’ from the spatial analyst tool box in ArcGIS 9.x was used to calculate the average distance to bikeways for each zip code. Here the zip codes were specified as ‘zones’.
c. The table output from the previous step was joined to the zip code shape file and the field indicating average distance was retained as the final field, Avg_d_bike
d. The unit of distance was converted to miles for consistency with the units of other variables included in this study.

11. Average Distance to Impaired lakes (Avg_d_imp)

The proximity to impaired lakes listed by MPCA (Avg_d_imp) (for details on ‘impaired lakes’ see section 2.4.2) in miles was calculated by the following steps:

a. Using the ‘Euclidean Distance’ function from the spatial analyst extension of ArcGIS 9.x, a Euclidean raster was created. This raster measured, for each cell size of 10 meter, the Euclidean distance to the closest source of impaired lakes.
b. ‘Zonal Statistics as Table’ from the spatial analyst tool box in ArcGIS 9.x was used to calculate the average distance to impaired lakes for each zip code. Here the zip codes were specified as ‘zones’.
c. The table output from the previous step was joined to the zip code shape file and the field indicating average distance was retained as the final field, Avg_d_imp
d. The unit of distance was converted to miles for consistency with the units of other variables.

Table 7 summarizes the proximity risk factors of WNV incidences.

Table 7 Sources and units of proximity risk factors of West Nile virus infection

Variables	Unit	Source
Avg_d_Type1	miles	MMCD
Avg_d_Type2	miles	MMCD
Avg_d_Type3	miles	MMCD
Avg_d_Type4	miles	MMCD
Avg_d_Type5	miles	MMCD
Avg_d_Type6	miles	MMCD
Avg_d_Type7	miles	MMCD
Avg_d_Type8	miles	MMCD
Avg_d_lakes	miles	MN DNR
Avg_d_openspace	miles	TLG, MetroGIS
Avg_d_sewers	miles	MetroGIS
Avg_d_waste	miles	MPCA, MCES
Avg_d_streams	miles	MetroGIS
Avg_d_golf	miles	TLG, MetroGIS
Avg_d_trails	miles	MC, MetroGIS
Avg_d_bike	miles	MnDOT, LMIC, MetroGIS
Avg_d_imp	miles	MPCA

2.4.4 Vector Control Factors

The arrival of WNV in Minnesota in 2002 elevated the importance of controlling *Culex* species mosquitoes. *Culex pipiens*, *Culex restuans*, *Culex salinarus*, and *Culex tarsalis* are the potential vectors carrying WNV in the TCMA. MMCD is the main agency responsible for surveillance and controlling of various vector species carrying infection for mosquito-borne diseases. The MMCD follows targeted chemical control programs, such as larviciding and adulticiding to control WNV vectors. Larva

control is the main focus of the program, but is supplemented by adult mosquito control measures when necessary (MMCD 2004).

MMCD records the larviciding and adulticiding applications by public land survey (PLS) units. Beginning in the late 1840s, the federal government began surveying Minnesota as part of the Public Land Survey System (PLSS). The resulting network of land survey lines divided the state into townships, ranges, sections, quarter sections, quarter-quarter sections and government lots, and therefore laid the groundwork for contemporary land ownership patterns. The source for this data set is the USGS 30-minute latitude by 60-minute longitude map series (1:100,000-scale). In Minnesota we can obtain this data for free from the LMIC. Based on the treatment data by PLS units, the following existing vector control variables were calculated for the year 2002 to 2007 at the zip code level:

1. **Percentage of PLS unit treated for both larva and adult mosquito control, (Perct_lar_ad)** at zip code (i) and time (j), where *i* ranges from 1 to 159 and *j* from the year 2002 to 2007

$$Perct\_lar\_ad_{ij} = \frac{PLS\ units\ treated\ for\ both\ Larvicide\ and\ Adulticide\ in\ zip\ code\ (i)}{Total\ PLS\ unit\ in\ zip\ code\ (i)} * 100$$

2. **Percentage of PLS unit treated for mosquito larva (Perct_lar)**

$$Perct\_lar_{ij} = \frac{PLS\ unit\ treated\ for\ larva\ in\ zip\ code\ (i)}{Total\ PLS\ unit\ in\ zip\ code\ (i)} * 100$$

3. **Percentage of PLS unit treated for mosquito larva (Perct_lar)**

$$Perct\_ad_{ij} = \frac{PLS\ unit\ treated\ for\ adult\ mosquitoes\ in\ zip\ code(i)}{Total\ PLS\ unit\ in\ zip\ code\ (i)} * 100$$

4. **Average Frequency of larviciding (Avg_Freq_lar)**

$$Avg\_Freq\_lar_{ij} = \frac{Sum\ of\ frequency\ of\ larviciding\ at\ all\ PLS\ unit\ in\ zip\ code(i)}{Total\ PLS\ unit\ treated\ for\ larva\ in\ zip\ code\ (i)}$$

5. **Average Frequency of adulticiding (Avg_Freq_lar)**

$$Avg\_Freq\_ad_{ij} = \frac{Sum\ of\ frequency\ of\ adulticidng\ at\ all\ PLS\ unit\ in\ zip\ code(i)}{Total\ PLS\ unit\ treated\ for\ adult\ mosquitoes\ in\ zip\ code\ (i)}$$

All these variables were processed spatially and calculated using ArcGIS 9.x and the 'Polygon in Polygon Analysis' function in Hawth's Tool extension (Hawthorne). Table 8 summarizes the existing vector control variables.

Table 8 Sources and units of West Nile virus vector control variables

Variables	Unit	Source
PLS Units		LMIC
Perct_lar_ad	%	MMCD
Perct_lar	%	MMCD
Perct_ad	%	MMCD
Avg_Freq_lar	Number	MMCD
Avg_Freq_ad	Number	MMCD

2.5 Summary

This chapter laid out the basic database created for the dissertation in terms of study area, WNV database, and the description of the data. The study area is the Twin Cities metropolitan area comprised of seven counties of Anoka, Hennepin, Carver, Scot, Dakota, Ramsey, and Washington. The virus first reached TCMA in 2002 and since then it has been a major concern for the MDH and MMCD. The WNV database is a personal geodatabase built in a GIS environment. Both WNV infected cases and risk factors were stored in the database and updated regularly as new data were obtained. The WNV incidences are comprised of bird, mosquito pool, and human cases. The potential risk factors are broadly divided into four categories: environmental (temperature, precipitation, vegetation, hydrologic features), built environment (catch basins, ditches, housing density, sewers), proximity (distance to lakes, bogs, swamps), and existing mosquito abatement programs. The sources, data type, and preprocessing techniques for WNV incidence data and the risk factors are described in details.

3. Chapter 3: Delineating West Nile virus transmission cycles at various scales as exposure areas: The nearest Neighbor Distance-Time model

3.1 Background

Various approaches are used to identify West Nile virus (WNV) exposure areas by using unusual sightings of *either* infected dead birds or human cases both prospectively and retrospectively. Delineating exposure areas are critical for public health officials to implement effective vector control measures and reduce the risk of infection in humans. In addition, exploring the exposure areas at local scale could aid in formulating hypotheses for further research. However, a significant and largely unmet need in WNV research is to incorporate the temporal characterization of virus spread and locational information of *three* components of transmission cycle, including birds (reservoir), mosquitoes (vector), and humans (host) on a localized scale. Exposure areas containing the main components of the WNV cycle in close proximity have higher potential to amplify an outbreak as compared to exposure areas delineated by a single component only. In this chapter, I introduce a novel approach, termed 'Nearest Neighbor Distance Time' or NNDT to delineate and retrospectively monitor WNV transmission cycles on various scales in the TCMA. The model is a combination of Geographic Information principles and ecological knowledge of WNV transmission. It uses the interaction of nonrandom spatial distribution of infected dead birds, positive mosquito pools, and human cases and the time required for the virus to be transmitted from one component to another in the cycle.

The chapter is organized as follows. This section critically reviews several other approaches used to identify WNV exposure and sets the need for a spatiotemporal model which incorporates information from all the three components of a transmission cycle. The section 3.2 describes the datasets used, section 3.3 details the NNDT methodology, section 3.4 reports the results of the NNDT technique and section 3.5 describes a set of computational controls to investigate the sensitivity of the technique in various scenarios. I conclude with a discussion of the implications of these results in the TCMA, mosquito abatement policies, and future lines of investigation in section 3.6.

Previous approaches to identify WNV exposure areas can be broadly divided into prospective and retrospective groups. Prospective techniques were primarily developed as early warning systems for WNV outbreaks (Mostashari et al. 2003b; Theophilides et al. 2003; Eidson et al. 2005; Corrigan et al. 2006; Johnson et al. 2006; Tachiiri et al. 2006; Theophilides et al. 2006). The initial and the most widely used methodologies to delineate WNV exposure areas appeared with the measurement of densities of WNV infected dead birds or measuring number of dead birds per unit area (Eidson, Karmer et al. 2001; Eidson, Miller et al. 2001; Mostashari et al. 2003a; Eidson et al. 2005). Typically density estimation results in a surface showing the spatial variation of local densities of points (location of WNV-infected dead birds) for a user specified search radius. In other words, a smoothly curved surface is fitted over each point, where the density value is highest at the location and diminishes with increasing distance from the point, reaching zero at the search radius or at the edge of the kernel for that point (Diggle 1985; Berman and Diggle 1989). Typically, regions with higher dead bird density are regarded as high risk areas and vice versa. In a similar approach, dead crow density (DCD) was calculated from real-time surveillance data for an areal unit, such as counties or tracts, and monitored at static time slices over a specific period of time (Eidson et al. 2005).

Even though these approaches produced quick and simple maps for identifying areas of high WNV activities they had number of limitations. First, density estimation algorithms selected an arbitrary cutoff (critical) radius to estimate density and identify

exposure areas for WNV, which had no association with the actual transmission of the virus. Second, density measurement was highly sensitive to the surveillance of infected dead birds reported by humans, which in turn were dependent on the confounding effects of population density, the varying level of interest of different communities in reporting dead birds, and absence of people in non accessible areas to report dead bird locations. Third, the assumption of uniform distribution of infected dead birds during the calculation of density did not hold true for spatial units varying in sizes and shapes, resulting in the aggregation issues of modifiable areal unit problem (MAUP) (Fotheringham and Wong 1991). Fourth, DCD calculated for large spatial units such as counties, neglected the local variation of densities or risk areas. Also the resultant high density areas based on counties will change if the base support was modified to tracts or blocks, resulting in the scale problems of MAUP (Fotheringham and Wong 1991) Fifth, density maps created using kernel functions showed inaccuracies of results near the boundaries due to edge effects (Theophilides et al. 2003). Sixth, the density measures were *only spatial* methods there by ignoring the dynamic temporal characterization of virus transmission. Studies using density algorithms only expressed time as static slices over a period. Seventh, they also did not consider the ecology of the WNV transmission cycle.

Another approach to identify WNV exposure areas for humans was to use the incidence cases of other components of the transmission cycle, namely infected dead birds (Theophilides et al. 2003; Johnson et al. 2006; Theophilides et al. 2006) or positive mosquito pools (Tachiiri et al. 2006), or equine cases (Corrigan et al. 2006) as proxies.. For example Theophilides et al. (2003) developed a Dynamic Continuous-Area Space-Time model (DYCAST) using the nonrandom spatiotemporal interaction of dead birds as an early warning for WNV infection in humans. The DYCAST system successfully identified areas of high risk for human WNV infection approximately 13 days prior to the onset of illness providing significant time for targeting remediation and control efforts. However, the efficiency of the DYCAST model can be increased further by incorporating the space-time interaction of mosquito pools as parameters, along with

infected bird cases.

In another study, a raster based mosquito abundance model for both the *Culex tarsalis* and the *Culex pipiens* was developed. The model produced maps for identifying WNV exposure areas for humans in British Columbia (Tachiiri et al. 2006). Similar to the DYCAST model, which included only the bird information, the raster based mosquito abundance model incorporated *only* the vector data, thus making the model weaker for predicting exposure areas for humans. A similar model associated significant spatiotemporal clusters of equine cases with the occurrences of human cases (Corrigan et al. 2006). Even though these studies significantly advanced several techniques to delineate exposure areas for human infection, they were based on only *one component* (dead birds or mosquito abundance or equine cases) of the transmission cycle. The models were incomplete given the multi-host nature of the infection. Therefore attempts to reduce the risk of WNV in humans based on such limited information probably are inadequate as well as costly.

In the retrospective group, the most common approach to delineate WNV exposure areas was the use of mathematical or statistical models. In these models the incidence cases of either infected dead birds, mosquitoes, or human cases were related to the predictor variables ranging from environmental, built environment and vector control policies (Brownstein et al. 2002; Ruiz et al. 2004; Bowman et al. 2005; Cooke, Katarzyna et al. 2006; Diuk-Wasser et al. 2006; Gibbs et al. 2006; David et al. 2007; Lian et al. 2007). Typically, in these models the WNV incidence was the response variable that was regressed by a set of predictor or WNV risk factor variables. Linear, logistic or Poisson link functions were used to indentify relationships between the predictors and the response variable. This resulted in the delineation of exposure areas based on the spatial distribution of important predictors for WNV incidences.

Although mathematical and statistical modeling attempted to overcome some of the limitations of density measures, they also had number of shortcomings. First, linear or logistic models assumed *a priori* that there was a linear relationship between the risk factors and WNV incidences. However, it is difficult for linear models to incorporate the

complexities of the WNV transmission network. As a result, previous investigations were not able to rigorously identify the exposure areas. Second, similar to the density-based approaches, these techniques typically center on a single host (dead birds or human cases) to identify WNV exposure areas, likely missing exposure areas involving all three components of the transmission cycle that have a higher potential to amplify a WNV outbreak as compared to exposure areas delineated by a single component only. Third, these regression based techniques did not incorporate the temporal characterization of virus transmission; they were only spatial models. Fourth, these models often aggregated the number of incidences and results to an arbitrary political unit, such as census tracts, zip codes, or blocks, leading to MAUP issues (Fotheringham and Wong 1991; Kulldroff and Hjalmars 1999; Theophilides et al. 2003).

The literature on methodologies that account for both spatial and temporal nature of virus transmission as well as incorporate multi-host ecological information of WNV to identify exposure areas retrospectively is rather scant. Orme-Zavaleta et al. (2006) applied a Bayesian probabilistic relational technique to generate spatiotemporal models using location and reporting dates of dead birds, positive mosquito pools, and human cases to investigate the nature of WNV transmission dynamics in Maryland (Orme-Zavaleta et al. 2006). Their probabilistic relational model served two purposes. First, it qualitatively and quantitatively investigated the spread of WNV and second, it acted as a tool to generate hypotheses related to the transmission of WNV in Maryland.

The aim of this study is to use Geographic Information Science (GIS) techniques and ecological knowledge of WNV transmission to develop a technique identifying exposure areas in their entirety involving birds, mosquitoes, and humans at various scales in the TCMA. The technique termed NNDT relies on space-time interaction among infected dead birds, positive mosquito pools, and human cases. The model can be useful to locate chronically exposed areas, formulate vector control policies, and allow researchers to generate hypotheses related to the transmission of WNV.

3.2 Data Description

The study used three datasets for the years 2002 to 2006. First, the Metropolitan Mosquito Control District (MMCD) provided the location and the date of WNV infected dead birds (American Crows and Blue Jays) reported by people. The raw data were later geocoded at the street level, verified, and cleaned with occasional field validation by MMCD staff. Second, MMCD also provided the location of WNV positive mosquito pools or traps designed to sample mosquitoes of different species for viral analysis on a weekly basis (MMCD 2004). MMCD has three types of collection traps for mosquitoes, which are evenly distributed in the TCMA. The trap types are CO2 and gravid traps, and sweep nets. The CO2 traps are elevated in the tree canopy and are used for collecting female mosquito samples in their host-seeking phase. The gravid traps located on the ground are designed to attract female mosquitoes that are seeking oviposition sites (i.e., places to lay eggs). As for the temporal coverage, mosquitoes are sampled for viral analysis on a weekly basis (MMCD 2004). Pool data included collection date, species type, and number of infected mosquitoes in the pool. Third, the Minnesota Department of Health (MDH) provided the location and the onset date of infected human cases by zip code. Because of confidentially of human health-related data and the guidelines of the University of Minnesota's Institutional Review Board (IRB) I obtained WNV-infected human cases aggregated by zip code. Table 9 details these data sets.

Table 9 Data Description

	2002	2003	2004	2005	2006	Spatial Scale	Temporal Scale	Attributes
Dead birds	102	285	125	60	479	Point	Day	X/Y, Date reported
Approx. No. of Mosquitoes (No. of Pools)	1100 (5)	1400 (17)	25 (2)	160 (14)	1550 (90)	Point	Day	X/Y, Collection date, Trap type, Species, No. of mosquitoes
Human Cases	13	26	6	7	15	Zip Code	Day	Onset date

3.3 Methodology

The 'Nearest Neighbor-Distance-Time' (NNDT) method evaluated distance thresholds based on the spatial distributions of WNV infected dead birds, positive mosquito pools, and zip codes with infected human cases in the TCMA. Hence, the estimated distance thresholds generated by this method were 'local' in the sense that they were sensitive to sub-regional variations in WNV incidence. Conducting NNDT involved: 1) calculating Euclidian distances to the nearest dead bird or mosquito pool location; 2) creating spider diagrams for these nearest neighbors centered on zip codes with human cases; 3) choosing spider webs connected to both the birds and the mosquito pools, 4) applying the temporal rule of virus transmission from one component to another, and 5) selecting spider webs for delineating transmission cycles. The following paragraphs describe each of these steps in detail.

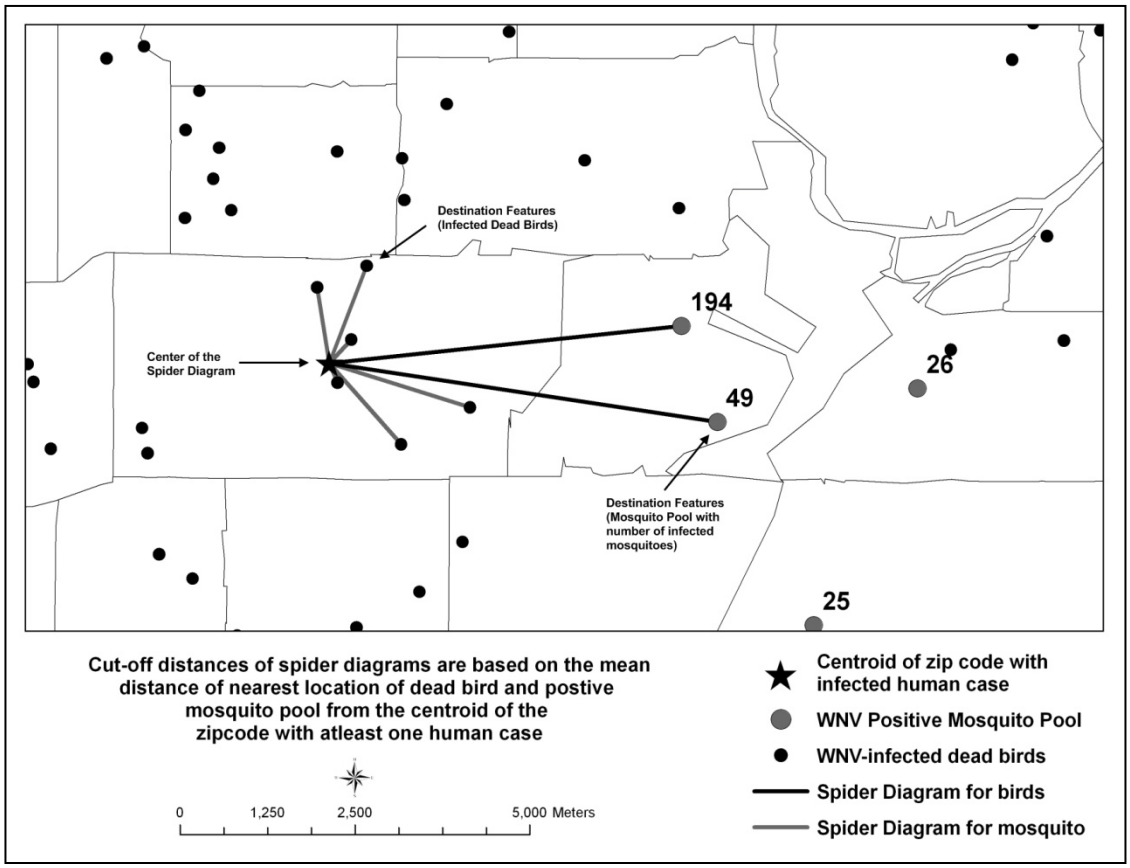
First, I calculated Euclidean distances from the centroid of zip codes with at least one WNV-infected human case to the nearest location of a dead bird and a positive mosquito pool (Table 10). The maximum, minimum, and the mean distances from the centroid of zip codes to the nearest location of a dead bird were highest in 2002. As mentioned in the previous chapters, WNV cases first appeared in 2002 in Minnesota and there were 102 reports of infected dead birds. This probably was an undercount because this was the first year that WNV was detected and there was lack of general knowledge or awareness of the disease early in the year. When incidences of WNV peaked in the year 2003, the distances dropped significantly, with the minimum and mean distances of location of infected dead birds from the centroid of zip codes were 0.1 and 1.2 miles respectively. There was an apparent negative relationship between the distance measures and the number of dead bird reports. For example, in the year 2005, there were only 60 locations of dead birds and the mean distance from the centroid of a zip code to a nearest location of a dead bird was as high as 4.7 miles. However, with the increase of dead birds reports in the year 2006 (479 cases), the mean distance decreased significantly to 1.8 miles. This is somewhat expected, given that all other things being equal, an increase in incidence also increases the likelihood that a human case will have a bird case nearby.

Table 10 Nearest neighbor Euclidean distances from the centroid of zip code with at least one WNV-infected human case (miles)

Year	Distance to Nearest Dead Birds			Distance to Nearest Positive Mosquito Pool		
	Maximum	Minimum	Mean	Maximum	Minimum	Mean
2002	14.9	0.93	4.35	21.4	1.5	11.8
2003	6.2	0.1	1.2	24.9	1.2	8.1
2004	6.2	0.4	3.1	23.6	1.0	13.7
2005	8.1	0.7	4.7	16.2	5.6	10.3
2006	11.8	0.9	1.8	11.2	1.1	3.7

It was a challenge to tease out spatiotemporal patterns from the Euclidean distances from the centroid of zip codes to the nearest positive mosquito pools mainly because of the nature of mosquito abatement and surveillance programs conducted by MMCD. The surveillance of *Culex pipiens*, *Culex restuans* and Culex *tarsalis* mosquito species, the main vectors of WNV transmission, increased from 2002 onward. Given that the traps are inspected weekly, temporal analysis of WNV positive mosquito pools will be biased in the sense they are not gathered daily. Nevertheless, the temporal resolution of a week can still be used to delineate WNV exposure areas.

Figure 13 An Example of Spider Diagram, 2006



In the second step, for each year, the mean distances from the centroid of zip codes to the nearest location of infected dead bird and mosquito pools were used as the cut-off distances for *spider* diagrams. A spider diagram joins a single source feature to a set of nearest similar destination features (points, lines, and polygons) based on a user specified distance threshold. Here the centers were the centroids of zip codes with at least one WNV human case and destination features were locations of dead birds and positive mosquito pools. Figure 13 shows an example of a spider diagram for the year 2006. Based on the mean distances, the center of the spider diagram connected six locations of dead birds and two mosquito pools by *spider legs*, and the formation of such a network including destination features, spider legs, and center or centroid, is called a *spider web*. Similar spider diagrams were generated for all zip codes with at least one human case from the year 2002 to 2006.

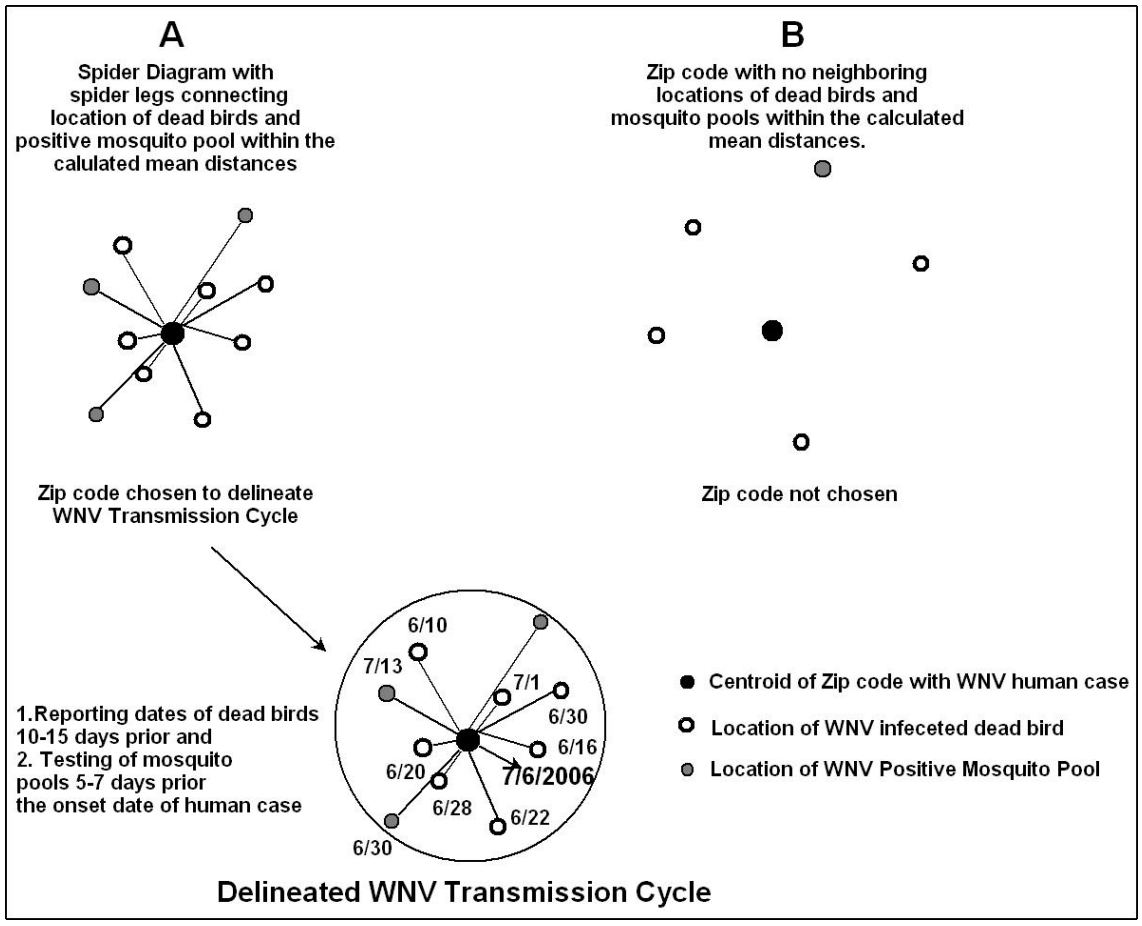
In the third step, the spider diagrams were categorized into 3 groups: spider diagrams connected to both, dead birds and mosquito pools, spider diagrams connected to locations of either birds or mosquito pools, and spider diagrams connected to neither. Only the first category of spider diagrams (those with both dead birds and pools) were selected for the next step.

In step four, the selected spider webs, centered on the centroids of zip codes were further classified by the temporal thresholds of transmission of the virus from birds to mosquitoes, and then from mosquitoes to humans. In the WNV transmission cycle, the virus is transmitted to humans by infected mosquitoes that acquire the virus by feeding on infected birds. Spider diagrams centered on zip codes were only selected when they satisfied the temporal sequence of actual WNV infections and when the reporting dates of the dead birds within the reach of the spider diagram (or*'web'*) were 10-15 days prior, and the testing dates of mosquito pools were 5-7 days before that of the onset date of human cases found in that zip code. These temporal thresholds are well documented components of the WNV disease cycle (Theophilides et al. 2003; Orme-Zavaleta et al. 2006; Theophilides et al. 2006). WNV research groups at the MMCD and the MDH also confirmed that these temporal windows were appropriate for

capturing bird-mosquito-human dynamics because infected birds generally tend to live 10 to 14 days after being infected by WNV. Mosquitoes are capable of transmitting the virus to another animal or humans within a week after feeding on infected bird (MMCD 2002; MDH 2005).

Finally, in step five, WNV transmission cycles were delineated by drawing circles centered on the zip codes of selected spider diagrams with a radius equivalent to the distance between the furthest destination feature (birds or mosquito pools) in that particular spider web. Figure 14 shows the steps involved in the NNDT method.

Figure 14 A Schematic Diagram of Nearest-Neighbor-Distance-Time (NNDT) methodology



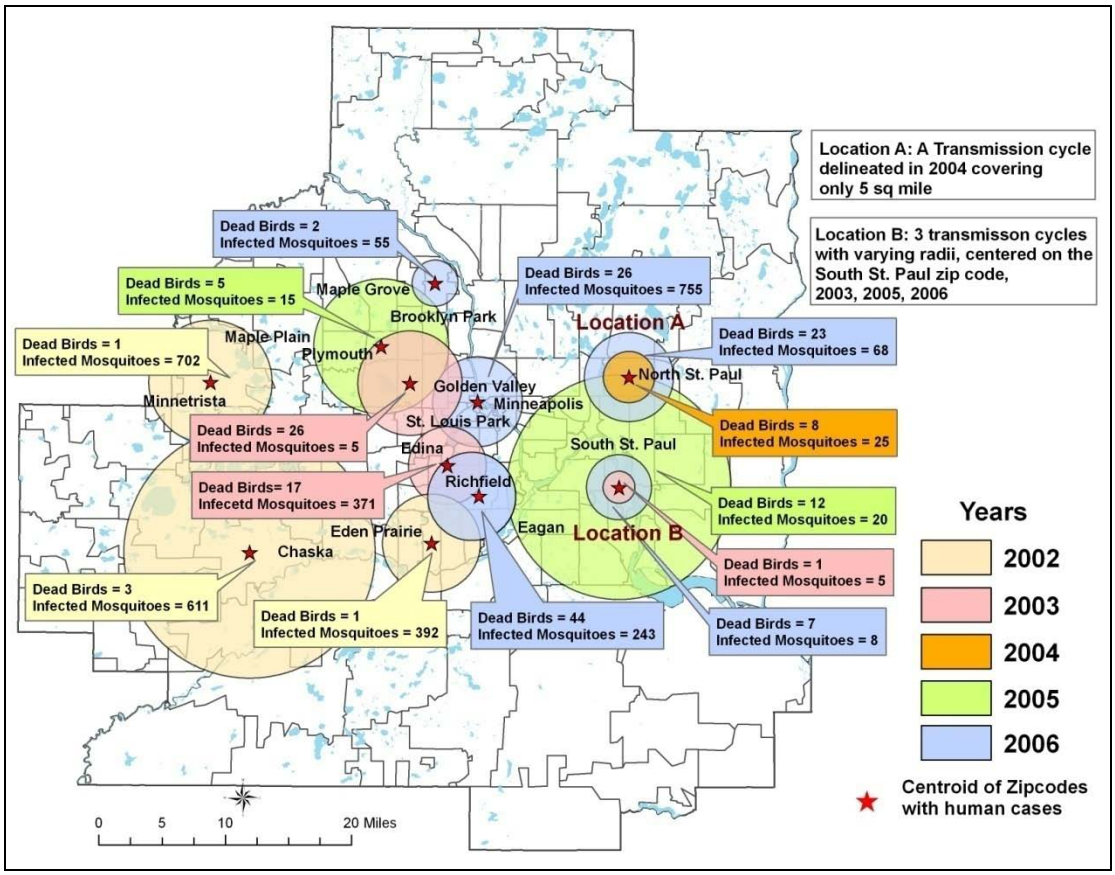
3.4 Results

The NNDT method retrospectively identified localized WNV transmission cycles with an area ranging from 5 to 150 sq miles. To summarize, NNDT calculated distance to the nearest dead bird or mosquito pool location, created spider diagrams for these nearest neighbors, and selected spider diagrams centered on zip codes by the extent to which they satisfy spatial and temporal rules used to delineate WNV transmission cycle. The years 2002 and 2003 had three WNV cycles each, followed by one and two in 2004 and 2005 respectively, and five cycles in 2006 (Table 11). I further plotted all the cycles on a single map to highlight the temporal trends (Figure 15).

Table 11 West Nile Virus incidences at a micro-scale, 2002-2006

Year	No. of Cycles	No. of infected dead birds	Percentage of Human Cases
2002	3	5	38
2003	3	44	27
2004	1	8	33
2005	2	17	57
2006	5	102	53

Figure 15 Location of delineated WNV transmission cycles at local scales, 2002-2006, Twin Cities Metropolitan Area. Minnesota



For the year 2002, three WNV transmission cycles were identified, all within 3 to 7 miles from the centroid of zip codes with positive human cases, and covering an aggregate area of about 255 sq miles. Per Figure 15, these cycles were located in the southwestern part of the metro area (Maple Plain, Minetrista, Chaska, and Eden Prairie). WNV peaked in the year 2003 along with increased bird surveillance and testing of mosquito pools. Three local WNV cycles were delineated within 1 to 3 miles from the zip codes with human cases. Compared to 2002, the 2003 cycles were much more intense and localized with 44 dead birds, 381 positive mosquitoes, and 7 human cases

within an area of approximately 90 sq miles mainly centered on the densely populated areas of Minneapolis, South St. Paul, Golden Valley, and Edina.

There were no spatial overlaps of transmission cycles in 2002 and 2003 because WNV activities shifted from the less-populated outlying suburban areas to the more densely populated urban areas in the center. There were two possible reasons for this spatial shift. The first reason could be attributed to the habitat characteristics of different mosquito species. According to the MMCD pool inspection archives, there was significant increase in *Culex tarsalis* mosquitoes in 2002, which were typically found in open prairie grasslands and dry short grasses that were common in the rural southwest metro. This increased vector population likely led to the formation of WNV transmission cycles in the southwest part of TCMA. Conversely, in 2003, there were more *Culex restuans* species, which were predominantly found in urban areas, exemplified by the delineation of WNV transmission cycles in Minneapolis and St. Paul. This increase of abundance of *Culex restuans* over *Culex tarsailis* resulted in the spatial shift of the transmission cycles. A similar outbreak of WNV occurred in the urban areas of Chicago in 2003 (Ruiz et al. 2004). The second reason lies with the surveillance system. By 2003, public understanding of WNV in Minnesota was widespread, in part due to the news in the local media of infections the previous year along with advertisements and precautionary measures issued by MDH. Greater awareness could create a higher rate of reporting of dead birds, which could increase the number of dead bird locations in areas of greater population (i.e., urban areas). In addition, MMCD also increased the frequency of testing of mosquito-breeding sites for WNV infection between 2002 and 2003.

Only one WNV transmission cycle was located in the year 2004. This particular cycle differed from previous cycles. The cycle was formed in a very small area of 5 sq miles, where the virus was transmitted from infected birds (8) to mosquitoes (25) and finally to humans (2) in North St. Paul (Location A in figure 4). The Euclidean distances from the human case (centroid of the zip code) to the nearest dead bird was approximately 0.35 miles and 1 mile to the WNV positive mosquito pool. Identification

of a transmission cycle in such a small area could be useful to investigate further and understand the potential causal factors of WNV infection. It would also allow researchers to investigate very detailed social and environmental characteristics that determine WNV incidence, such as land use, population characteristics, presence of stagnant water features, or mosquito abatement programs.

The WNV cycles identified in 2005 and 2006 shared several similarities with those of 2003. The two 2005 cycles were similar in that they were in urban areas, with 17 dead birds, 35 infected mosquitoes and 4 human cases within a combined area of approximately 200 sq miles. In 2006, five WNV transmission cycles had a total area of about 130 sq miles and were similar to the 2003 cycles in that they were smaller and more intense than other years. The Euclidean distances of dead birds and positive mosquito pools from the human cases located at the centroid of the zip codes were all within a close proximity of 1 to 3 miles and had relatively large combined figures of 102 infected dead birds, 1129 infected mosquitoes, and 8 human cases. The 2006 cycles, like those of 2003 and 2004, were in urban areas of Brooklyn Park, Minneapolis, Richfield, and South St. Paul.

Trend analysis of the WNV transmission cycles from 2002 to 2006 indicated the following characteristics. In spatial terms, after 2002, the location of the cycles shifted from the outer western suburbs and southwestern TCMA area to densely populated areas near the core cities of Minneapolis and St. Paul. Of these areas, South St. Paul was identified as exposure areas for the years 2003, 2005, and 2006 (Location B in figure 4), and therefore emerged as an important location for further investigation to investigate the causal factors of WNV activities.

In terms of temporal dynamics, reporting of dead birds after the onset of a human case dropped significantly i.e., on average the number of dead birds reports declined from 10 to 12 cases before, to one or two cases after the human onset date. This was likely because of localized reduction in bird population due to either bird die-offs or migration, leaving mosquitoes to feed on dead-end hosts like humans and other mammals, especially horses. There was also evidence that the WNV cycle reached its

peak 10-15 days prior to the onset of human illness and by the time humans developed WNV symptoms the transmission of virus were disrupted in that area due to lack of substantial number of infected birds.

3.5 Sensitivity Analysis

The preceding discussion highlighted the methodology and the spatiotemporal interpretation of results obtained from the NNDT model. I performed a series of computational controls to investigate the behavior of the NNDT technique in different scenarios, primarily focusing on randomization tests and an examination of the assumptions as a form of model validation.

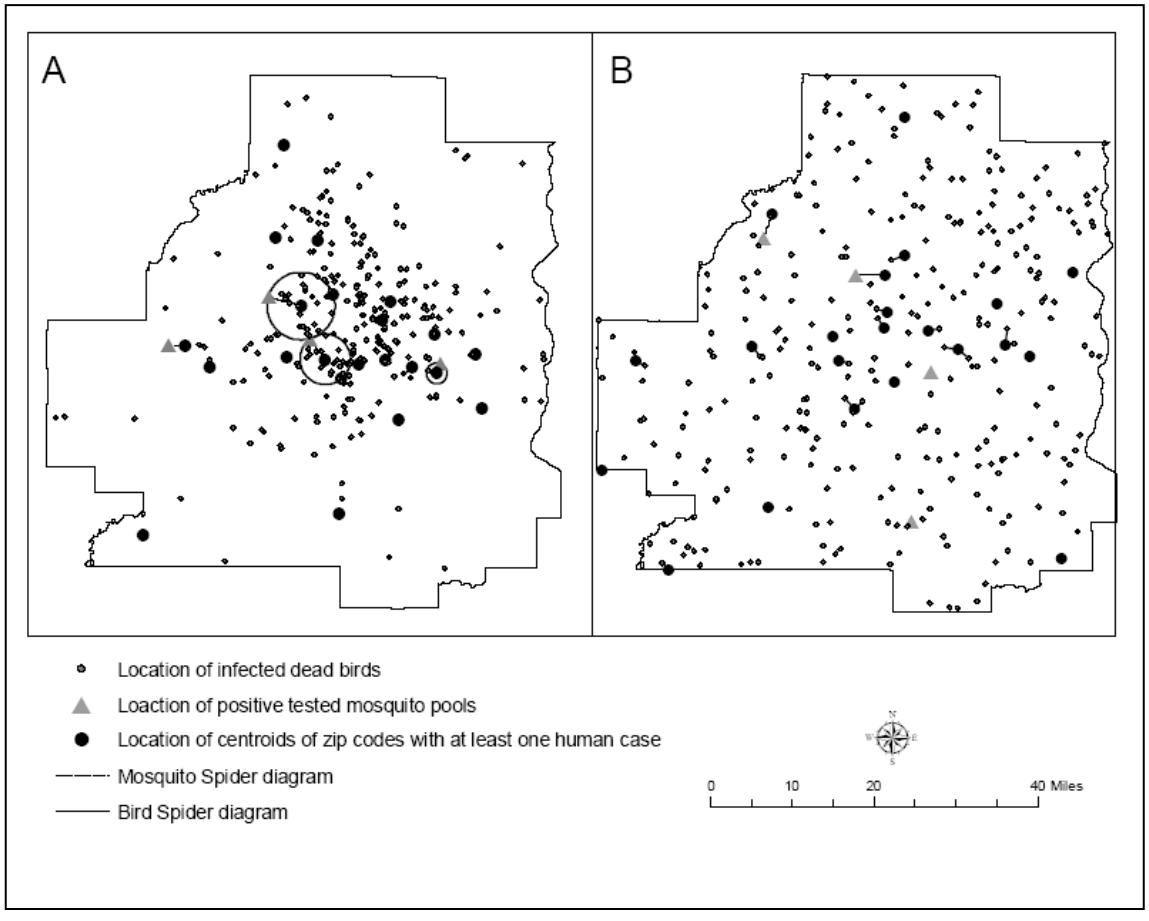
3.5.1 Randomization Test

Randomization tests by shuffling the locations and the reporting dates of WNV infected dead birds, positive mosquito pools, and human cases reported at the centroids of the zip codes were conducted. In both the cases, it is expected to lose any kind of pattern in the spatial distribution of each of the components, correlation between the proximity of one component to another, and the temporal sequence in reporting of WNV infected dead birds, mosquitoes, and humans. Given that NNDT is primarily based on the nonrandom spatial and temporal distribution of the three main components, it should not identify any WNV transmission cycles from the random data.

First I randomized the locations of dead birds, mosquitoes, and human cases while keeping their reporting dates the same. A dataset was generated by randomly choosing the locations for each of the components while keeping the number of incidences same for a particular year. For example, the generated dataset for the year 2003 had 285 random locations of infected dead birds, 4 WNV positive mosquito pools

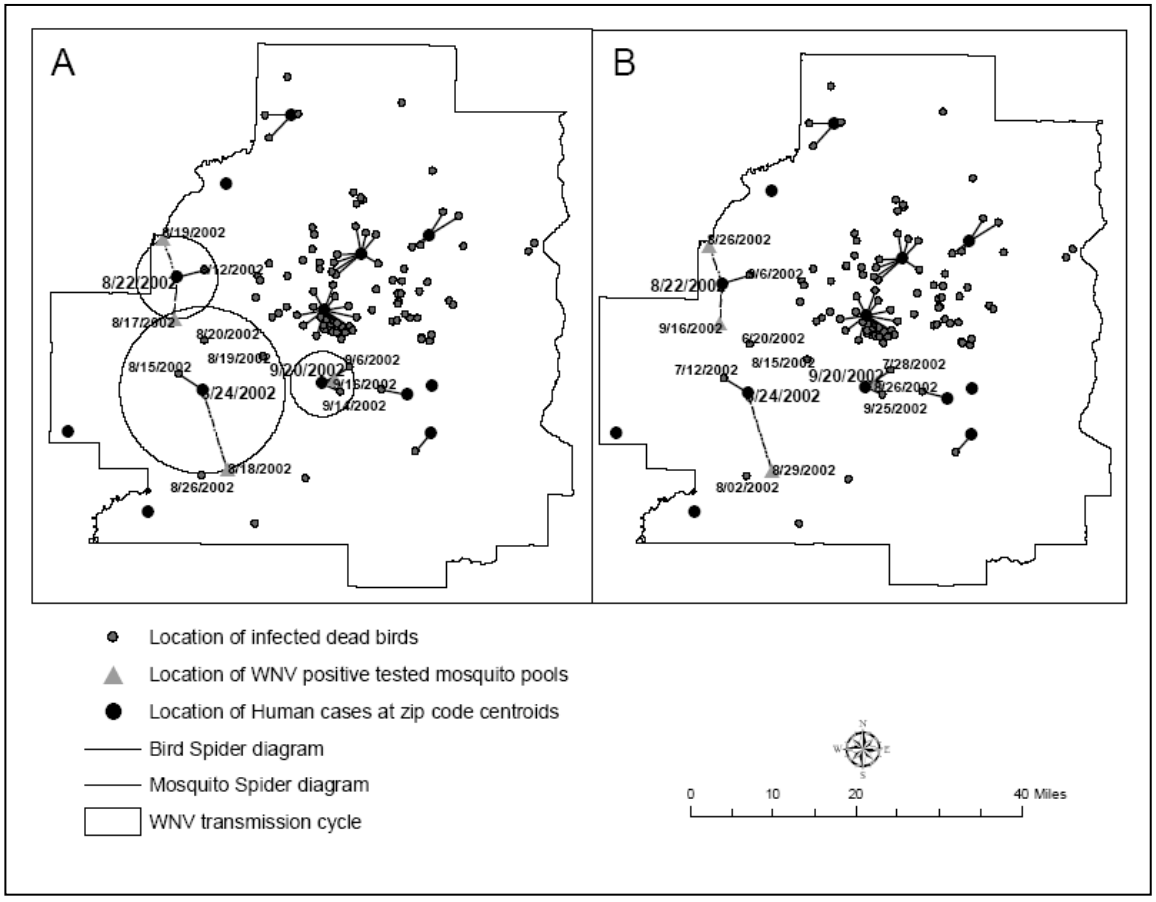
and 22 infected human cases. Figure 16 presents a comparison of the application of NNDT methodology between a real dataset and a random dataset for the year 2003. Part A represents the results of NNDT on the real data, from which the technique delineated 3 WNV transmission cycles. The Part B shows the results obtained from the noisy data. Here the NNDT methodology halted at step 3, where the spider diagrams centered on centroids of zip codes with at least one human case did not connect to both the locations of dead birds and the mosquito pools based on the distance thresholds computed in step 1. Hence no transmission cycles were indentified for this particular dataset.

Figure 16 Comparison of the application of NNDT methodology between a real dataset and a noisy dataset for the year 2003



Similar results were obtained from other datasets for the years 2002, 2004, 2005, and 2006. The NNDT either failed at step 3 or at step 4, where by temporal rules were applied to the selected centroids joining both the locations of birds and the mosquito pools. In this step, the zip codes were further selected if in a particular spider web, the reporting dates of dead birds were 10-15 days prior and the mosquito pools were 5-7 days earlier than the reporting dates of human cases assigned to the centroid of zip codes. Comparing the results to the real datasets, it appeared that chance plays little part in identifying WNV transmission cycles by the NNDT technique.

Figure 17 Comparison of NNDT when applied to a real and a "time" scrambled data set for the year 2002



Next, I tested the temporal aspect of the NNDT model by randomizing the reporting dates of dead birds, infected mosquito pools, and human cases while keeping their spatial locations the same. As expected, the NNDT will not detect any WNV transmission cycles from the scrambled datasets because the temporal sequence of virus transmission from birds to mosquitoes and then to humans will be lost. Figure 17 shows the outcome of NNDT when applied to a real and a randomized data set for the year 2002. Based on the spatial rules only, the NNDT technique identified 3 spider centers or centroids of zip codes connecting to the destination features of dead birds and mosquito pools for both the real (Part A) and the scrambled dataset (Part B). As shown in part A, all the three centroids fulfilled the temporal rules, i.e. almost 90 percent of the birds in the spider web were reported 10-15 days prior and mosquito pools were reported positive for WNV 5-7 days earlier than the reporting dates of human incidence at that zip code. Hence we can delineate WNV transmission cycles centered on those centroids. However the figure in Part B shows that even when we have same centroids selected, the failure of temporal rules did not allow the formation of any WNV transmission cycles.

Finally, both the locations of birds, mosquito pools, and human cases were randomized, as well as their reporting dates. The NNDT approach did not find any transmission cycles. The methodology either did not find a centroid of zip code connecting both the locations of dead birds and mosquito pools or did not satisfy the temporal rules. These tests indicate that the NNDT methodology is robust and the delineation of transmission cycles based on spatial and temporal rules did not arise by chance in the case examined here.

3.5.2 Modification of the assumptions

There are two key assumptions of the NNDT model. First, due to human health data confidentiality reasons, I only obtained the number of WNV infected human cases aggregated at the zip code level. Thus, to calculate the distance between the locations of a human case to the nearest location of an infected dead bird, I assumed that the human

cases were reported at the centroid of the zip codes. Second, to compute the cut-off distances to draw the spider diagrams, I also assumed that the *mean* of the distances between the zip codes and the nearest location of dead birds and mosquitoes pools was an appropriate distance indicator. The following sections describe scenarios for which these assumptions were modified.

3.5.2.1 Varying the locations of human cases

To investigate the sensitivity of NNDT method, the calculations with human cases reported at several random locations within a particular zip code were repeated. This was carried out by running the NNDT technique ten times, for each year from 2002 to 2006.

I first calculated the average distances to the nearest location of infected dead birds and mosquito pools from the location of human cases aggregated at random points in the zip codes. Further, I used these distances to draw spider diagrams for birds and mosquito pools from the respective random points. The results of distance calculations are summarized in Table 12. Except for the year 2005, the range of mean distances to the nearest location of dead birds and mosquito pools computed from the random locations of human cases were within approximately 1 mile. The averages computed for the year 2006 has the smallest range of 0.62 miles and 0.63 miles to the locations of dead bird and mosquito pool respectively. Figure 18 depicts the error plot of the distances computed from the ten random location of human cases and the nearest location of a dead bird and a mosquito pool in 2006, where each vertical lines represents one nearest neighbor distance calculation with maximum (top cap), mean (solid dot), and minimum (bottom cap) values. The solid and dashed horizontal lines are the average, maximum and minimum values of distances calculated from the centroid of the zip code. For all the iterations of the NNDT technique with random locations, the mean distances were very similar to the value calculated from the centroid of zip codes. This was true for both the birds and mosquito pool locations. Such small changes in the averages when calculated from the random points compared to the average calculated

from the centroid of the zip code have little or no effect on the size of spider diagrams and the resultant delineation of transmission cycles.

Table 12 Comparison of statistics of mean distance thresholds computed from random locations of human cases and centroid of zip codes to the locations of infected dead birds (Bird) and positive mosquito pools (Mosq)

	2002		2003		2004		2005		2006	
Run	**Bird**	**Mosq**	**Bird**	**Mosq**	**Bird**	**Mosq**	**Bird**	**Mosq**	**Bird**	**Mosq**
R1	4.35	11.18	1.86	8.70	3.42	13.67	4.97	10.56	2.49	4.35
R2	4.35	11.18	1.24	8.08	3.42	14.91	4.97	10.56	1.86	4.35
R3	3.73	11.81	1.24	8.08	2.49	13.67	4.35	11.18	2.49	3.73
R4	3.73	11.81	1.86	8.39	2.49	13.05	5.59	10.56	2.49	3.73
R5	4.35	11.18	1.24	8.08	3.11	14.29	6.21	9.94	2.49	4.04
R6	4.97	11.81	1.55	8.08	3.73	14.29	4.66	11.18	1.86	3.73
R7	4.35	11.81	1.86	8.08	3.42	14.91	4.97	9.01	2.17	3.73
R8	4.35	11.81	1.24	8.08	3.42	13.67	3.42	9.94	1.86	3.73
R9	4.97	11.18	1.55	7.46	3.11	13.05	4.35	9.32	1.86	3.73
R10	4.35	11.81	1.55	8.39	2.49	13.67	4.97	10.56	2.17	4.04
Centroid	**4.35**	**11.81**	**1.24**	**8.08**	**3.11**	**13.67**	**4.66**	**10.25**	**1.86**	**3.73**
Range	1.24	0.62	0.62	1.24	1.24	1.86	2.80	2.17	0.62	0.63

However, for the year 2005, the range of average distances to the nearest location of infected dead birds was 2.8 miles and for the mosquito pools was 2.2 miles (Table 12). Figure 19 shows two error bar plots for the distances calculated from 10 random locations of human cases to dead bird and mosquito pool locations in 2005. For the dead birds, only two out of 10 averages were within a difference of a mile from the average distance calculated from the centroid of zip codes. Conversely, for the mosquito

pools six averages were within a mile from the actual average distance. Such variation in the average distances calculated from the random points will likely influence the size and location of spider diagrams and eventually the delineation of WNV transmission cycles. Comparing the results of 2005 to other years, I found two plausible reasons for such wide variation. First, there were far fewer incidences of infected dead birds, positive mosquito pools, and human cases as compared to other years. Second, the zip codes reported with at least one WNV infected human case in 2005 were bigger in area than the zip codes reported human cases for other years. Hence it can hypothesized that the NNDT methodology is sensitive to the number and density of infected cases but performs well with higher number of cases. The technique will have better results if the aggregating units for the human cases are smaller in size.

Figure 18 The error plot of the distances computed from the 10 random location of human case and the nearest location of a dead bird and a mosquito pool in 2006

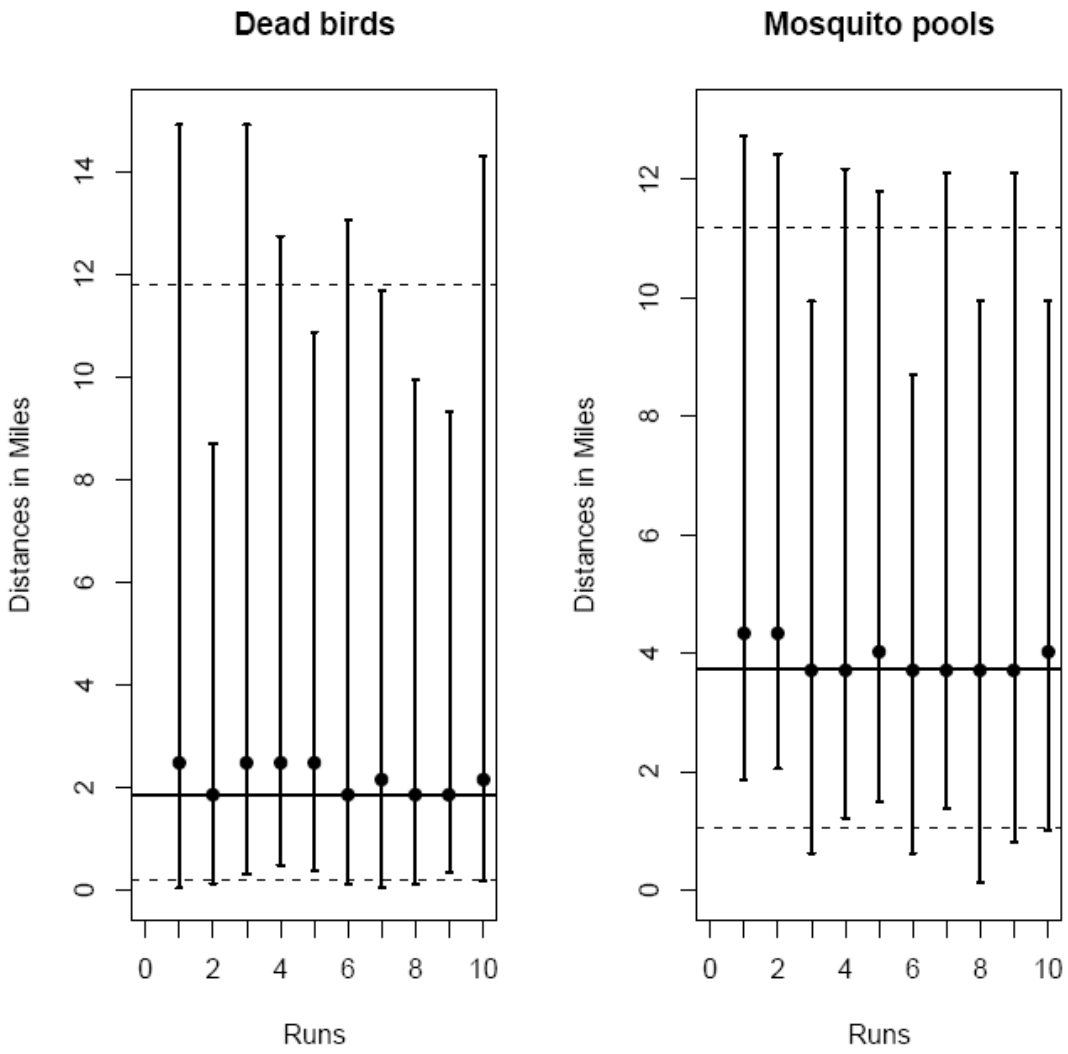
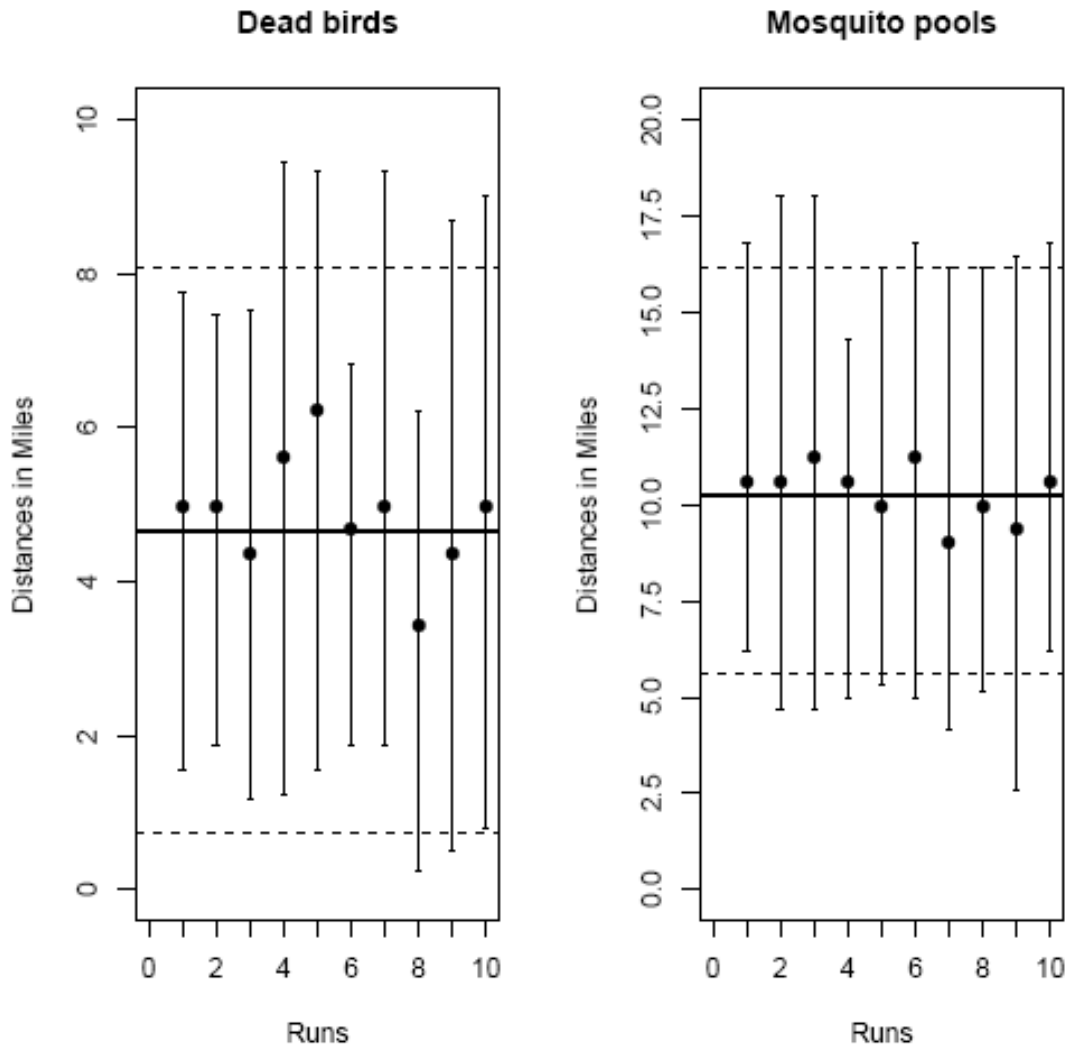


Figure 19 The error bar plots for the distances calculated from 10 random locations of human cases to dead bird and mosquito pool locations in 2005



3.5.2.2 Varying the distance indicator

I also analyzed the results of NNDT when using the maximum and minimum distances between the centroid of zip codes and the location of infected dead birds and positive mosquito pools, and compare the resultant findings to those obtained with mean distances.

When using the minimum distance criteria between the location of infected dead birds, positive mosquito pools, and human cases located at the centroid of the reported zip codes, the NNDT identified no transmission cycles. This result was consistent across all years in that none of the centroids or spider diagram centers connected both the destination features (birds and mosquitoes) and hence halted the NNDT methodology at step 2.

Using the maximum distance as the cut-off to draw spider diagrams yielded different results as when compared to the use of mean distance. First, almost all the delineated WNV transmission cycle increased in size. For example, for the years, 2004, 2005, and 2006, the exposure areas increased in size, while in 2002 and 2003, areas of two transmission cycles remained the same. Second, the length of the spider legs for the bird and mosquito spider diagram increased, thereby connecting more dead birds and mosquito pools located further away from the centroid of the zip code with at least one infected human case. This did not necessarily result in either more transmission cycles or larger ones, primarily because the reporting dates of these new destination features (birds and mosquito pools) did not satisfy the temporal conditions (fell outside of the 10 to 15 days for dead birds and 5 to 7 days for mosquito pools). Thus if we exclude the destination features which did not satisfy the temporal rules, the structure of the spider diagrams and the resulting WNV transmission cycles resembled to those delineated using the mean distances. Third, using the maximum distance as a threshold resulted in more centroids of zip codes fulfilling the spatial rule of NNDT by connecting to both the locations of infected dead birds and positive mosquito pools in 2004, 2005, and 2006. However, these new centroids did not meet the temporal criteria for reporting dates. In sum, using a minimum distance threshold did not identify any transmission cycles, while using the maximum distances resulted in more spider webs with both dead birds and mosquito pools but did not satisfy the temporal criteria of the NNDT model.

3.6 Discussion and Conclusion

This chapter demonstrated the use of NNDT as a new approach to delineate WNV exposure areas. NNDT uses GIS techniques and ecology of the virus to retrospectively delineate transmission cycles as exposure areas in their entirety, including dead birds, mosquito pools, and human cases. The NNDT methodology is based on the combination of distance-time interaction and the temporal sequence of virus transmission from one component to another. As such, this approach improves upon the use of arbitrary "critical" density measures, goes beyond the use of clusters of a single component of the WNV transmission cycle, and provides more flexibility in some respects than statistical models based on *a priori* relationships (although it does not address multivariate causality in the way statistical models do). More broadly, NNDT demonstrates how examining distances among the locations of infected dead birds, positive mosquito pools, and infected human cases infuses a 'local' dimension that is combined with the 'global' knowledge of the time required for the virus to be transmitted from one component to another in the cycle. NNDT is therefore a retrospective quantitative methodology in which the relationship between the distances of spatial locations and the temporal thresholds of the transmission of virus from one component to another within a cycle is used to delineate exposure areas.

NNDT results provide evidence that dead bird reports are an essential part of the delineation of WNV cycles at localized scale. In the TCMA, the cycle peaked 10-15 days prior to the onset of human illness, a period that was consistent with the epidemiology of the virus transmission from bird to mosquito and then to humans. Temporal analysis of the delineated cycles at local scales also showed evidence of substantial reduction in the reporting of dead birds after the 10 -15 days window either due to localized reduction in bird population or migration or both. This indicated a signal of reduction in amplification activities due to lack of abundant reservoir (bird) population and virus proliferation ceased further with the onset of human illness. These findings can influence remediation and control strategies. For example, preventive strategies such as insecticide spraying may be best concentrated during the temporal

window of 10-15 days after the reporting dates of WNV-infected dead birds because by the time humans develop WNV symptoms in an given area, the risk of further amplification of WNV disease may have subsided and thus widespread spraying of insecticide will have little or no effect. More efficient allocation of resources allows for a better balance between the need for mosquito eradication and desire to limit the environmental impacts from insecticide usage.

Delineation of WNV cycles at local scales can also be used to develop hypotheses related to the disease spread and thus advance our understanding of the complexity of avian-mosquito-human environmental systems on a micro scale. The NNDT methodology, for example, delineated a WNV transmission cycle in North St. Paul in 2004. The cycle remained active for 12-16 days prior to the date of human illness and then subsided with substantial decrease in dead bird reporting. These dynamics point to the need to investigate the feeding behavior of mosquitoes and their habitat systems. The delineation of WNV cycles and their components at local scale helps such localized investigations by prompting questions such as “Why did a WNV cycle occur here” or “What environmental and neighborhood characteristics catalyze the formation of WNV transmission cycles?”

Despite the success of the NNDT methodology, certain issues require further research and are worthy of mention. The NNDT technique is sensitive to the number of incidences as noted above for the year 2005. It performs more efficiently with higher number of infected dead birds or mosquito pools or human cases, and as such is dependent upon the efforts of surveillance programs. Also, the technique will have better results if the human cases are obtained at much finer spatial resolution than the zip codes.

This chapter mainly concentrated on developing a novel technique to retrospectively visualize and delineate WNV transmission cycles in its entirety at various scale. A critical future research direction includes investigation of the complex avian-mosquito-human and environmental interrelationships that create WNV high-risk exposure areas. Another area of exploration understands the extent to which NNDT can

be used to understand WNV in other areas of the world given that the characteristics of transmission cycle and surveillance programs vary significantly from one region to another.

4. Chapter 4: Association of potential risk factors and West Nile Virus illness in the Twin Cities Metropolitan Area of Minnesota

4.1 Background

The rate and intensity of WNV transmission in birds, mosquitoes, and humans are also influenced by different natural and man-made conditions. Exploring the association of such natural and anthropogenic *risk factors* or triggers can aid in understanding the occurrence of WNV incidences. This would allow identifying risk factor(s) whose management would result in effective disease prevention and containment. I selected several hypothesized risk factors based on thorough review of previous studies, discussions with experts at MDH and MMCD, and my domain knowledge of the study area. The risk factors were broadly divided into four categories: environmental, proximity, built-environment variables, and mosquito abatement policies. The Chapter 2 provides a description of each of the risk factors with data sources, formats, and preprocessing methods.

The focus of this chapter is to explore the association of potential risk factors associated with WNV disease in birds, mosquitoes, and humans in the TCMA. The chapter is organized as follows. The sub sections of 4.1 discuss the selection of risk factors in respect to previous studies, section 4.2 details the data and methodology used, section 4.3 reports the results, section 4.4 provides a discussion, and finally section 4.5 concludes the chapter.

4.1.1 Environmental Factors

Several environmental factors affecting habitat characteristics of both avian hosts and vector population play important role in the transmission of WNV. Key environmental risk factors include vegetation cover (Brownstein et al. 2002; O'Leary et al. 2004; Ruiz et al. 2004; Hayes et al. 2005; Cooke, Grala et al. 2006; Gibbs et al. 2006; Warner et al. 2006; Ezenwa et al. 2007; Leblond et al. 2007; Ruiz et al. 2007; DeGroote et al. 2008; Ozdenerol et al. 2008), physiographic factors (Rappole et al. 2000; Cooke, Katarzyna et al. 2006; Ruiz et al. 2007; DeGroote et al. 2008; Ozdenerol et al. 2008), hydrologic features (Rappole et al. 2000; Cooke, Katarzyna et al. 2006; Ezenwa et al. 2007; Leblond et al. 2007; DeGroote et al. 2008), and weather conditions (Theophilides et al. 2003; O'Leary et al. 2004; Ruiz et al. 2004; Hayes et al. 2005; Gibbs et al. 2006; Savage et al. 2006; Theophilides et al. 2006; Vaidyanathan and Scott 2006; Adlouni et al. 2007; Bolling et al. 2007; Gleiser et al. 2007; Landesman et al. 2007; Zou et al. 2007; Bouden et al. 2008; DeGroote et al. 2008; Ozdenerol et al. 2008).

One of the earlier studies based in New York, revealed a positive correlation between spatial distribution of WNV incidents and vegetation abundance (Brownstein et al. 2002). In Chicago, during the 2002 outbreak, similar association between WNV activities and vegetation cover was observed (Ruiz et al. 2004). Ezenwa *et al's* investigation of risk factors in Louisiana indicated that wetland areas were positively correlated with WNV infected dead birds locations but were negatively correlated with the infection rates of *Culex* mosquitoes (Ezenwa et al. 2007). In Mississippi among other variables, stream density, slope, and Normalized Difference Vegetation Index (NDVI) contributed to the prediction of WNV disease incidences at the zip code level (Cooke, Grala et al. 2006).

Temperature and precipitation play the most important role in the breeding of susceptible vector population that further increases the risk of virus circulation between birds and mosquitoes. Potential mosquito carriers then transmit the virus to humans through their bites. As mentioned above, there are several studies, which investigated the effect of weather conditions and the occurrence of WNV illness. For example, Bell

et al. concluded that in Red River Valley of North Dakota, cooler weather conditions in 2004 were the principle reason for lower number of WNV cases (Bell et al. 2006). Degree day (DD) models developed from temperature and precipitation data were also used to determine when and where WNV reached a sufficient vector infectivity level (Zou et al. 2007). A similar study concluded that the 2002 WNV outbreak in the Northeastern American cities was characterized by two main variables: 1) lower number of DD below -5C in the winter, and 2) higher number of DD greater than 25C in the summer (Adlouni et al. 2007). A national study at the county level found that the virus occurrence in a particular year was strongly associated with annual precipitation from the previous year (Landesman et al. 2007). Shaman *et al.* associated the occurrence of drought 2 – 6 months prior and land surface wetting 0.5 – 1.5 months before with the WNV infection in sentinel chickens and humans in Florida. The authors explained that the drought brought avian hosts and mosquitoes into close contact and thus facilitated the transmission and amplification of the virus (Shaman et al. 2005).

4.1.2 Socioeconomic, Demographic, and Built-Environment Factors

Socioeconomic and demographic risk factors include age, gender, population density, and income (O'Leary et al. 2004; Ruiz et al. 2004; Hayes et al. 2005; Gibbs et al. 2006; Warner et al. 2006; Hayes 2007). Even though all ages appear susceptible to WNV infection, a study showed that risk was especially higher for people among 60 to 80 years of age, and was slightly higher among male patients (Hayes et al. 2005). In Chicago, during the 2002 outbreak, WNV infected human cases were clustered in areas with predominance of older white Caucasian residence with medium to low population density (Ruiz et al. 2004). In Georgia, Gibbs *et al.* found that the rate of WNV infection increased in the populated urban/suburban areas and decreased in the mountainous regions of the state (Gibbs et al. 2006).

Built-environment variables or man-made factors also influence the variability of virus transmission from mosquitoes to humans. The urban features that are hypothesized with WNV illness are housing density (Ruiz et al. 2004; Gibbs et al. 2006;

Hinckley et al. 2007; Ruiz et al. 2007), age of house (Ruiz et al. 2004; Ruiz et al. 2007), road density (Cooke, Katarzyna et al. 2006; DeGroote et al. 2008), sewage treatment lagoons (Savage H. and Miller 1995; Huhn et al. 2005; Hayes 2007), storm water ponds and urban catch basins (Savage and Miller 1995; MMCD 2004; Huhn et al. 2005; Rey et al. 2006; Stockwell et al. 2006), construction sites that retain standing water (O'Leary et al. 2004), stagnant water in scrap-tire stock piles (*Florida Statutes* 2002), waste water discharge sites (Zou et al. 2006), and roadside ditches (Savage and Miller 1995; Huhn et al. 2005). In addition, I included unpaved trails and bikeways as favorable habitats for vector populations.

The study conducted by Ruiz *et al.* in Chicago indicated that census tracts with higher housing density and houses built between 1950 to 1960 increased the likelihood to include at least one WNV infected human case (Ruiz et al. 2004). In Georgia the urban/suburban areas with medium housing density had higher probabilities of being associated with WNV infected birds (Gibbs et al. 2006).

As mentioned in the previous chapters, *Culex* species is the most important vector population responsible for transmitting and spreading the virus in the United States. Thus knowledge about their feeding behavior and spatial variability of their habitat characteristics is essential to investigate the spread of WNV. *Culex* mosquitoes deposit their eggs in stagnant waters, which can accumulate in low areas with poor drainage, urban catch basins, storm water ponds, ditches, puddles, sewage treatment lagoons, construction sites, stock piles of old tires, and water discharge sites. Rey *et al.* analyzed how salinity, type of structure, and setting of catch basins and storm water ponds influenced the breeding and abundance of *Culex* mosquitoes in Florida (Rey et al. 2006). The Powder River basin in north central Wyoming experienced higher risk of WNV infection due to increase in potential larval habitats from 1999 to 2004. This area saw significant increase in coalbed methane gas extractions since the late 1990s. As a result large volumes of water were discharged creating additional aquatic habitats for mosquito breeding (Zou et al. 2006).

4.1.3 Proximity Factors

Nearness or *proximity* of risk factors also plays an important role in the spread of WNV. Several studies indicated that nearness to hydrologic features namely wetlands, streams, estuaries, lakes (Rappole et al. 2000), WNV infected dead birds (Brownstein et al. 2002; Theophilides et al. 2003; Ruiz et al. 2004; Eidson et al. 2005; Johnson et al. 2006; Koenig et al. 2007; LaDeau et al. 2007; Shaman 2007), infected mosquito pools (Hayes et al. 2005; Warner et al. 2006; Zou et al. 2006; Shaman 2007), and man-made aquatic habitats (storm-water ponds and water discharge sites) (Zou et al. 2006) increased the likelihood of WNV infection in humans.

4.1.4 Mosquito Abatement Policies

The temporal and spatial coverage of existing mosquito abatement policies in a region will affect the abundance of mosquito population. This in turn will influence the rate of WNV transmission from mosquitoes to humans (Ruiz et al. 2004; Hayes et al. 2005). Thus considering the abatement policies as an influencing factor in the amplification of the virus is important. In Chicago a census tract's assignment to either of the Mosquito Abatement Districts (MAD) or the City of Chicago or Dupage County, emerged to be an significant factor in understanding the spatial pattern of disease occurrence (Ruiz et al. 2004). The study indicated the possibility of differential risk to WNV from living in one MAD verses another.

4.2 Data and Methodology

Data on the WNV infected dead birds, mosquitoes, and human cases were aggregated at the zip code level. Spatially explicit data on each of the risk factors were also analyzed by zip codes (refer chapter 2 for the preprocessing of the data). The risk

factors in each of the four categories of environmental, built-environment, proximity, and vector abatement policies are described in Table 13.

Statistical analyses were carried out using R statistical software. The steps involved in the methodology are as follows:

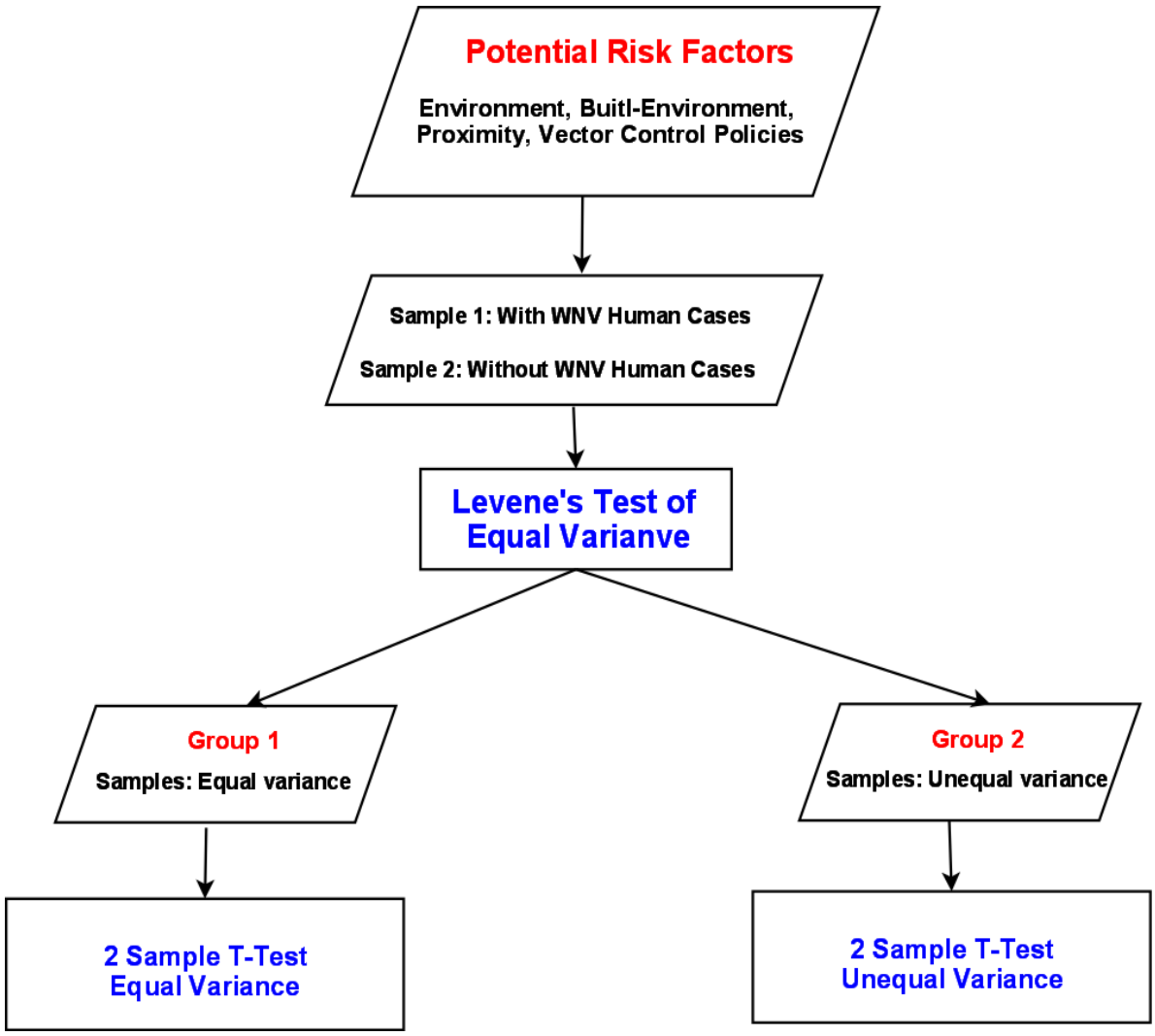
1. Divided each risk factor into two independent samples:
    a. Sample 1: Value of the risk factor for zip codes with WNV infected human case
    b. Sample 2: Value of the risk factor zip codes with no human case
2. Conducted *Levene's test* of equal variance. Levene's test is used to test if *k* samples have equal variances. Equal variance across samples is called homogeneity of variances. Some statistical tests, for example the analysis of variance or t-tests, assume that variances are equal across groups or samples. The Levene's test can be used to verify that assumption. It tests the null hypothesis (H0:) that the population variances are equal. If the p-value of Levene's test is less than a specified critical value (here 0.05), the differences in sample variances are unlikely to have occurred due to chance. In such a situation the null hypothesis of equal variances is rejected and it is concluded that there is a difference between the variances in the population.
3. Based on the results of Levene's test, the samples are sorted into two groups:
    a. Group 1: Sample which fulfilled the assumption of equal variance (p-value > 0.05) at 95 percent confidence interval
    b. Group 2: Sample which violated the assumption of equal variance (p-value < 0.05) at 95 percent confidence interval
4. Conducted two independent sample t-test with equal variance assumed for the samples in Group 1 and two independent sample t-test with unequal variance assumed for the samples in Group 2. This was done to test whether mean values of environment, built-environment, proximity, and vector control risk factors for zip codes with and without human WNV cases were statistically different.

Figure 20 describes the methodology in a flow diagram.

Table 13 Potential Risk Factors of West Nile virus illness in the Twin Cities Metropolitan Area of Minnesota

Categories	Factors
Environment	Maximum Daily Temperature, Minimum Daily Temperature, Daily Precipitation, Area of 8 types of Wetlands, Area of Lakes, Area of Open Green Space, Land Cover (14 classes), Average Elevation, Density of Streams/sq. mile, Average Dew Point
Built-Environment	Density of urban catch basins/sq. mile (dry), Density of urban catch basins/sq. mile (wet), Density of Ditches/sq. mile, Area of Impaired Lakes, Density of Sewer/sq. miles, Housing density/acre, Age of houses, Density of Roads/sq. mile, Density of Bike path/ sq. mile, Density of Population
Proximity	Distance to 8 types of Wetlands, Distance to WNV infected mosquito pools, Distance to lakes, Distance to Open Green Space, Distance to sewers, Distance to waste water discharge points, Distance to Streams, Distance to Golf Courses, Distance to Trails, Distance to Bike paths, Distance to Impaired lakes
Vector Control Policies	Percentage of public land survey(PLS) units treated for both adult mosquito and larva in 2002, Percentage of PLS units treated for both adult mosquito and larva in 2006, Percentage of PLS treated for larva 2002, Percentage of PLS treated for adult mosquitoes 2002, Average Frequency of Larvicide Treatment in 2002, Average Frequency of Adulticide Treatment in 2002, Percentage of PLS treated for larva 2006, Percentage of PLS treated for adult mosquitoes 2006, Average Frequency of Larvicide Treatment in 2006, Average Frequency of Adulticide Treatment in 2006

Figure 20 Flow Diagram showing the methodology to investigate the association between the potential risk factors and West Nile virus disease incidence



4.3 Results

The sample size of zip codes with infected human cases was 130 and that of zip codes without human cases was 29. Based on this all the above mentioned risk factors were divided into two samples of zip code with and without human cases. The Levene's test of equal variance further divided the samples into two groups of equal and unequal variance (Table 14). The environmental variables with unequal variances are maximum daily temperature, minimum daily temperature, wetland types of inland fresh marsh, inland fresh open water, and bogs, and land cover variables including open water, developed low density, developed high density, barren land, evergreen forest, pasture/hay, and woody wetlands. The remaining environmental factors failed to reject the null hypothesis (H0: Equal Variance) at 95 percent significance level indicating that their variances were equal.

The built-environment factors in Group 2 (unequal variance) were density of catch basins (wet), area of impaired lakes, density of sewer lines, and housing density. The remaining factors, density of catch basins (dry), density of ditches, road density, density of bikeways, population density, and age of houses were in Group 1 (Table 14).

Among the proximity risk factors, only distance to lakes had equal variance. Among the vector control policies, percentage of treated public land survey (PLS) units for both larva and adult mosquitoes (2006), percentage of treated PLS unit separately for larva and adult mosquitoes (2002, and 2006), and average frequency of adulticide treatment (2002) were in Group 2. The remaining policy variables belonged to Group 1 with equal variance.

The following paragraphs report the t-test results.

Table 14 Results of Levene's Test of Equal Variance

Levene's Test of Equal Variance	
Group 1 (Equal Variance) p-value >0.05, 95% level of Significance	**Group 2 (Unequal Variance) p-value <0.05, 95% level of Significance**
Precipitation	Maximum Daily Temperature
Seasonally Flooded Basins and Flats	Minimum daily Temperature
Inland Fresh Meadow	Inland Shallow Fresh Marsh
Inland Deep Fresh Marsh	Inland Fresh Open Water
Shrub Swamps	Bog
Wooded Swamp	LC_11_perct_area
Percentage of Open Green Space	LC_22_perct_area
LC_21_perct_area	LC_24_perct_area
LC_23_perct_area	LC_31_perct_area
LC_41_perct_area	LC_42_perct_area
LC_43_perct_area	LC_81_perct_area
LC_52_perct_area	LC_90_perct_area
LC_82_perct_area	
LC_95_perct_area	Density of Catch Basins (Wet)
Elevation	Area of Impaired lakes
Stream Density	Density of Sewers
Average Dew Point	Housing Density
Density of Catch Basins (Dry)	Perct_treat_pls_lar_ad_06
Density of Ditches	Perct_lar_treated_pls_02
Road Density	Perct_ad_treat_pls_02
Density o f Bikeways	Perct_lar_treat_pls_06
Population Density	Perct_ad_treat_pls_06
Average Age of Houses	Avg_freq_ad_02
Perct_treat_pls_lar_ad_02	Distance to infected mosquito pool
Avg_freq_lar_02	Distance to Open Green Space
Avg_freq_lar_06	Distance to waste-water discharge sites
Avg_freq_ad_06	Distance to Golf courses
	Distance to Trails
Distance to Lakes	Distance to Bikeway paths
	Distance to Impaired lakes

Note: Environmental, Built-Environment, Vector Control Policies, and Proximity Factors

4.3.1 Environmental Factors

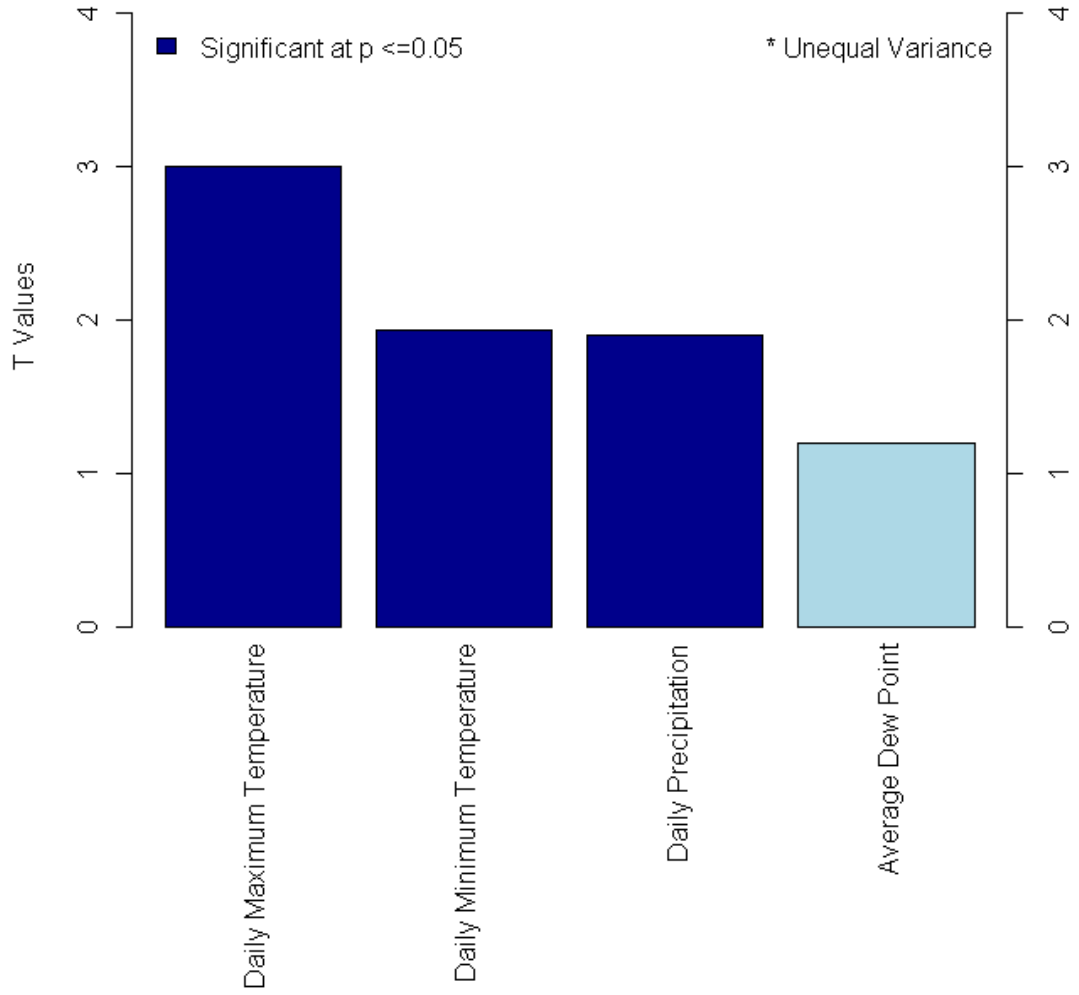
The t-test results for the environmental factors are reported in Table 15 and Figure 21**.** There was a statistically significant difference (t-value = 3.003, p-value = 0.00453) in mean daily maximum temperature between zip codes with (N = 130) and without WNV infected human cases (N = 29). Figure 21 graphically represents the t-values on the y axis and the factors on the x axis. The mean of daily minimum temperature (61F) and daily precipitation (0.036 inches) were higher in zip codes with WNV incidences than zip codes with none. However average dew point temperature did not show significant differences (t-value = 1.2024, p-value = 0.2351) between the two groups of zip codes.

Table 15 T-test results for Weather Factors

Variables	t	P-value	95% Lower CI	95% Upper CI
Daily Maximum Temperature	**3.003**	**0.004537**	**-2.58714**	**-1.42842**
Daily Minimum Temperature	**1.9317**	**0.05487**	**-1.80286**	**-1.87563**
Daily Precipitation	**1.8985**	**0.05346**	**-0.04502**	**-0.04850**
Average Dew Point	1.2024	0.2351	-3.33943	12.83594

Note: The mean of factors in **bold** are significantly different in zip code with human West Nile virus cases than zip codes without human cases at 95 percent significance level.

Figure 21 T-values for weather variables based on Two Independent Sample t-tests



The t-values resulting from the Student T-test (equal variance) and Levene's T-test (unequal variance) of land cover variables in zip codes with or without WNV disease incidence is presented in Table 16 and in Figure 22. Further, Figure 23 represents the land cover of TCMA with the location of WNV infected dead birds, positive mosquito pools, and centroids of zip codes with human cases. Means of two very distinct types of land cover classes emerged as statistically significant. Significantly higher means of developed open space, developed low density, developed medium density, and developed high density were found in zip codes with human disease cases. This indicated a positive association between developed urban spaces and occurrence of WNV illness. On the other hand, land cover classes such as pasture/hay (t-value = -2.3969) and cultivated crops (t-value = -2.0785) showed significantly negative association with disease occurrence. Land cover class of woody wetlands also had higher mean value in zip codes with a human case than zip codes with no cases.

Table 16 T-test results for Land Cover Factors

Factors	t	p-value	95% Lower CI	95% Upper CI
Developed, Low Density	**4.3533**	**0.000002409**	**6.65410**	**17.70763**
Evergreen Forest	**3.5138**	**0.0005774**	**0.41711**	**1.48806**
Developed, High Density	**3.0998**	**0.002296**	**-13.89586**	**-3.07924**
Developed, Open Space	**3.003**	**0.004537**	**1.15862**	**5.91645**
Woody Wetlands	**2.3279**	**0.02119**	**0.04300**	**0.52468**
Developed, Medium Density	**1.875**	**0.04802**	**-0.32564**	**-0.72836**
Open water	0.9816	0.3278	-1.47041	4.37577
Deciduous Forest	0.8274	0.4129	-1.91694	4.57651
Barren Land	0.7396	0.4606	-0.32770	0.72002
Mixed Forest	0.5462	0.5882	-0.04600	0.07996
Emergent Herbaceous Wetlands	0.4498	0.6551	-1.64177	2.58515
Shrub/Scrub	-1.3317	0.1913	1.14664	0.23764
Cultivated Crops	**-2.0785**	**0.04487**	**-18.52908**	**-0.22676**
Pasture/Hay	**-2.3969**	**0.01771**	**-9.89932**	**-0.95475**

Note: The mean of factors in **bold** are significantly different in zip code with human West Nile virus cases than zip codes without human cases at 95 percent significance level.

Figure 22 T-values for land cover variables based on Two Independent Sample t-tests

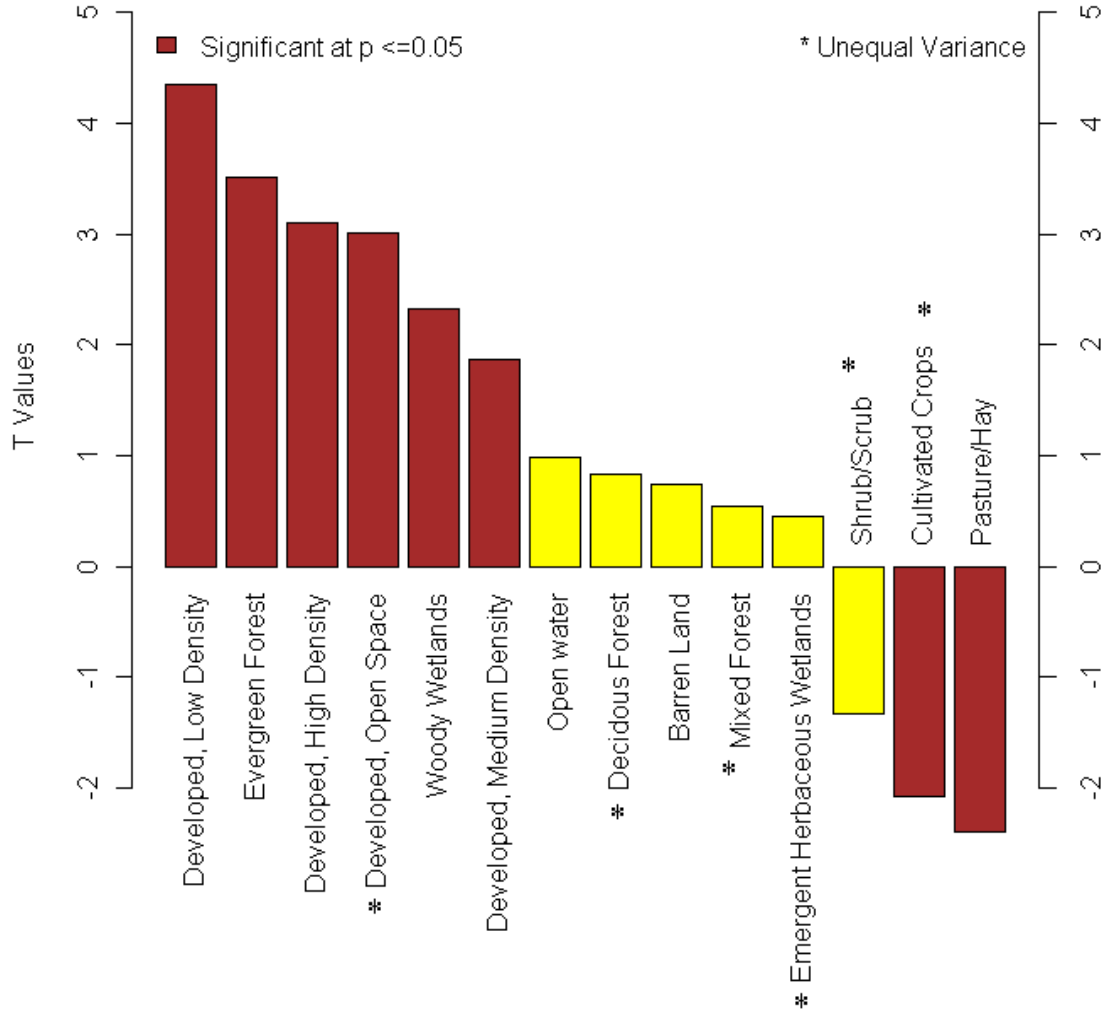
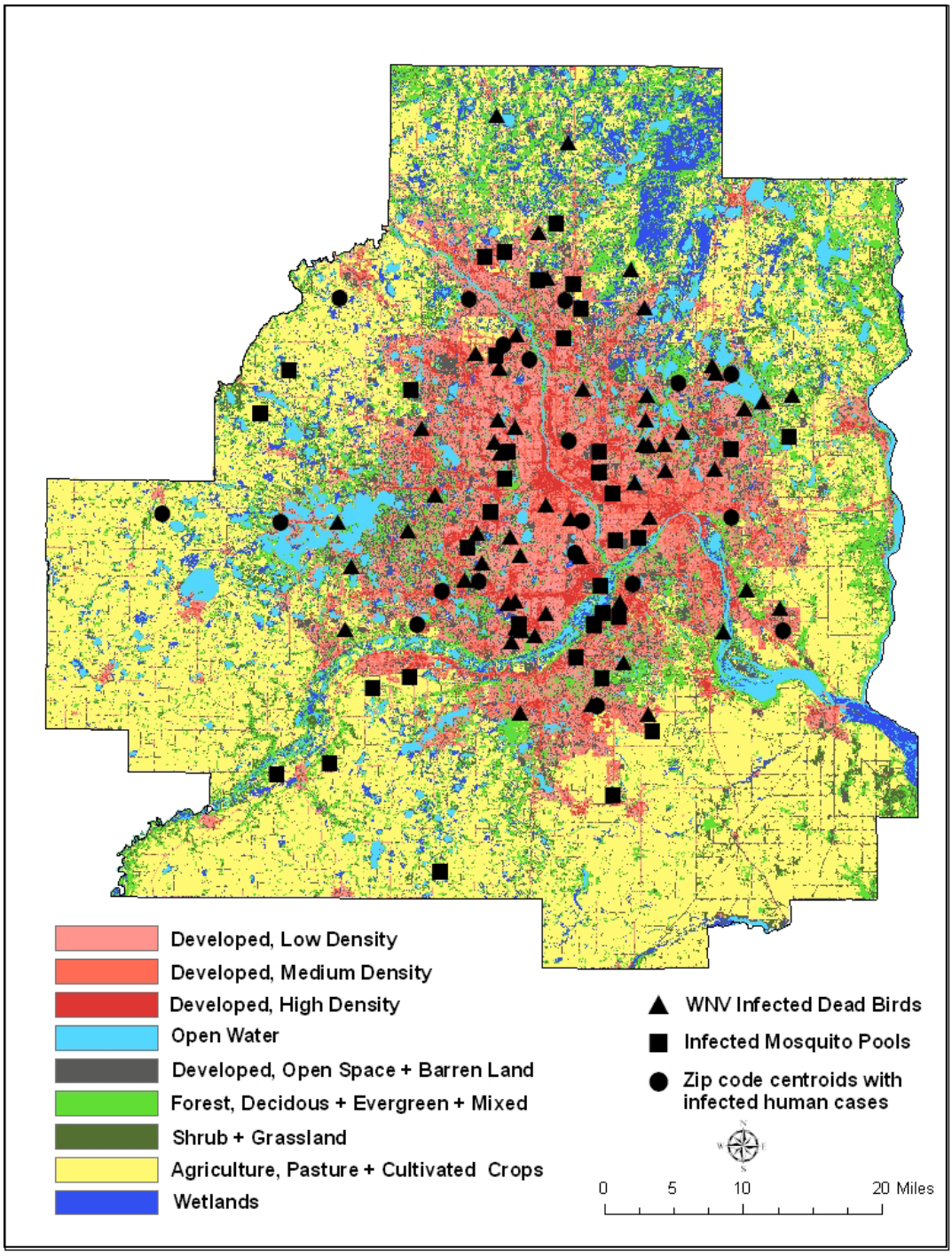


Figure 23 West Nile virus infected dead birds, positive mosquito pools, and centroids of zip codes with human cases overlain on TCMA 2001 land cover



Among the different wetlands, distance to five out of eight types showed statistically significant negative association with WNV disease occurrence (Table 17 and Figure 24). The five wetland types in descending order of t-values were distance to bogs (t-value = -3.64), distance to inland fresh open water (t-value = -3.417), distance to shrub swamps (t-value = -1.9985), inland shallow fresh marsh (t-value = -1.8985), and distance to wooded swamps (t-value = -1.898) (Figure 24). They had statistically significant lower mean values of distances to wetlands (bogs = 6284.54 meters, inland fresh open water = 1408.02 meters, shrub swamps = 2874.78 meters, inland shallow fresh marsh = 529.20 meters, and wooded swamp = 5920.68 meters) in zip codes with infected human cases than zip codes without disease incidents. Figure 25 shows the negative spatial association between one such wetland type (bogs) and WNV activities.

Table 17 T-test results for other Environmental Factors

Factors	t	p-value	95% Lower CI	95% Upper CI
Bog (D)	**-3.64**	**0.0003696**	**-6013.52600**	**-1782.90600**
Inland Fresh Open Water (D)	**-3.4717**	**0.0006682**	**-1675.98650**	**-460.47110**
Shrub Swamps (D)	**-1.9985**	**0.05768**	**-2343.72850**	**-882.00390**
Inland Shallow Fresh Marsh(D)	**-1.8985**	**0.05946**	**-293.04691**	**-5.80038**
Wooded Swamp (D)	**-1.898**	**0.05568**	**-3096.72800**	**-2479.34900**
Stream Density	-0.7716	0.4452	-0.38312	0.17171
Elevation	-0.7269	0.4713	-3.768796	1.772379
Inland Deep Fresh Marsh (D)	-0.7114	0.4814	-419.64000	201.71470
Seasonally Flooded Basins and Flats (D)	-0.5238	0.6032	-192.83640	113.37260
Inland Fresh Meadow (D)	-0.1405	0.889	-216.58770	188.48370
Percentage of Open Green Space	0.9306	0.358	-1.77894	4.80398

Note: The mean of factors in **bold** are significantly different in zip code with human West Nile virus cases than zip codes without human cases at 95 percent significance level. The variables with (D) represent the distance to a risk factor, for e.g., distance to bogs, distance to wooded swamp, etc.

Figure 24 T-values for other environmental variables based on Two Independent Sample t-tests

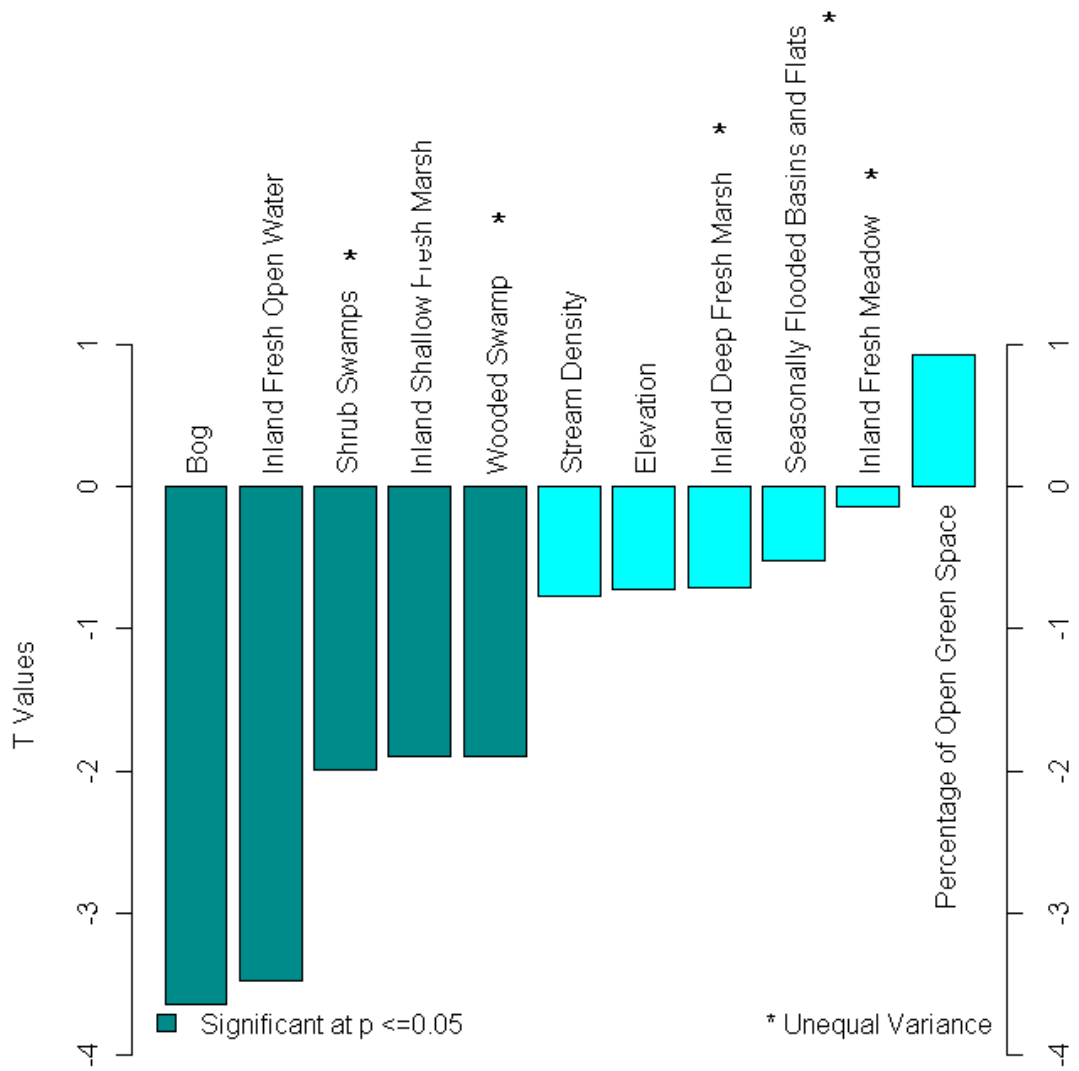
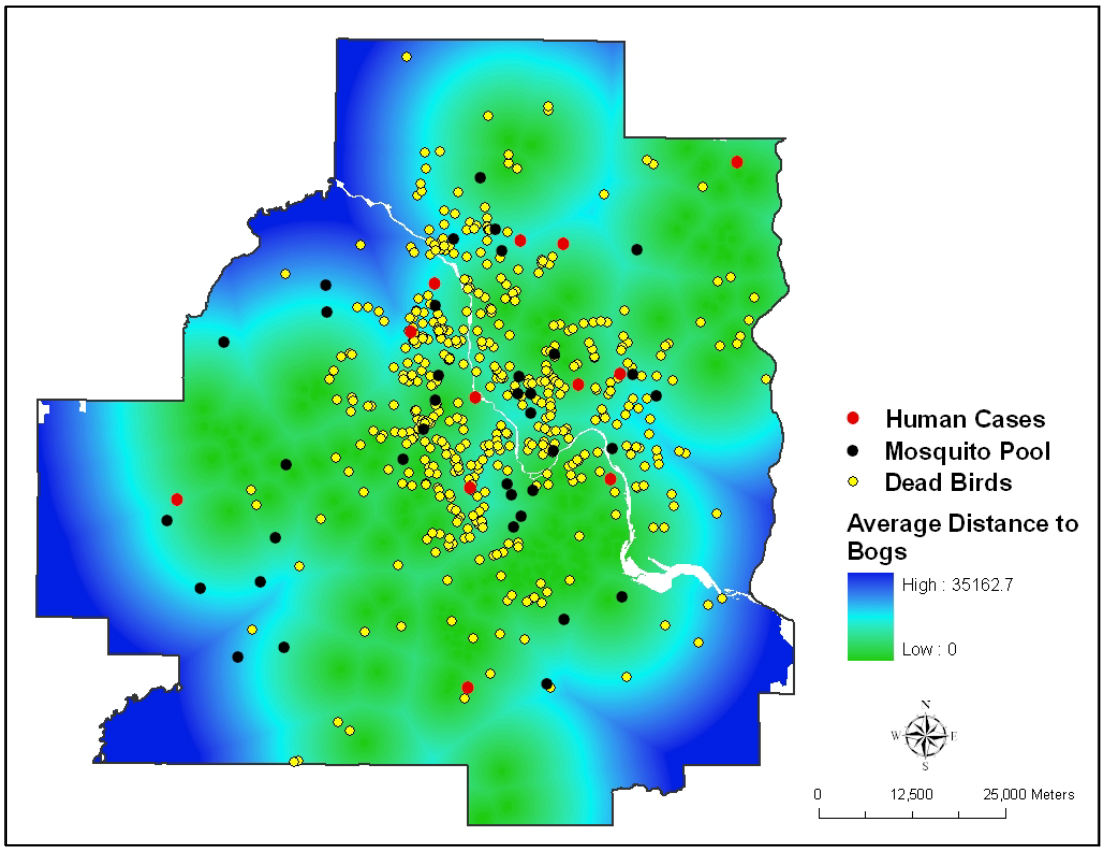


Figure 25 An overlay of West Nile virus incidences in 2006 on average distance to bogs (meters)



4.3.2 Built-Environment Factors

The t-test for means of built-environment factors in zip codes with and without human WNV disease incidence is presented in Table 18 and Figure 26. There was a statistically significant difference (t-value = 3.7917, p-value = 0.0002) in mean housing density between zip codes with (N = 130, mean = 2.035/acre) and without WNV human disease incidences (N = 29, mean = 0.373/acre).

The Figure 27 shows the spatial distribution of density of buildings (commercial and residential) along with the location of WNV incidences. Age of houses also had significant higher mean values (45 years) in zip codes with human cases. Another important finding was that the mean density of urban catch basins (wet) was statistically higher in zip codes with human cases. The other built-environment variables, namely density of catch basins (dry), density of bikeways, density of roads, percentage of area of impaired lakes, density of sewers, and density of ditches did not show statistically significant difference in mean values between zip code with and without WNV infected human cases.

Table 18 T-test results for Built-environmental Factors

Factors	t	p-value	95% Lower CI	95% Upper CI
Housing Density	**3.7917**	**0.000213**	**0.31706**	**1.00659**
Age of Houses	**2.0058**	**0.0487**	**-6.49362**	**-1.14986**
Density of Catch Basins (Wet)	**1.99712**	**0.05142**	**-6.98205**	**-1.49643**
Density of Catch Basins (Dry)	1.29	0.2045	-30.17900	136.60940
Density of Bikeways	1.2064	0.2352	-0.21316	0.84147
Density of Roads	1.1731	0.2485	-1.48267	5.55134
Percentage of area of Impaired Lakes	0.9316	0.353	-1.27979	3.56488
Density of Sewers	0.4264	0.6704	-0.20170	0.31275
Density of Ditches	-0.0818	0.9352	-0.09988	0.09212

Note: The mean of factors in **bold** are significantly different in zip codes with human West Nile virus cases than zip codes without human cases at 95 percent significance level.

Figure 26 T-values for built-environment variables based on Two Independent Sample t-tests

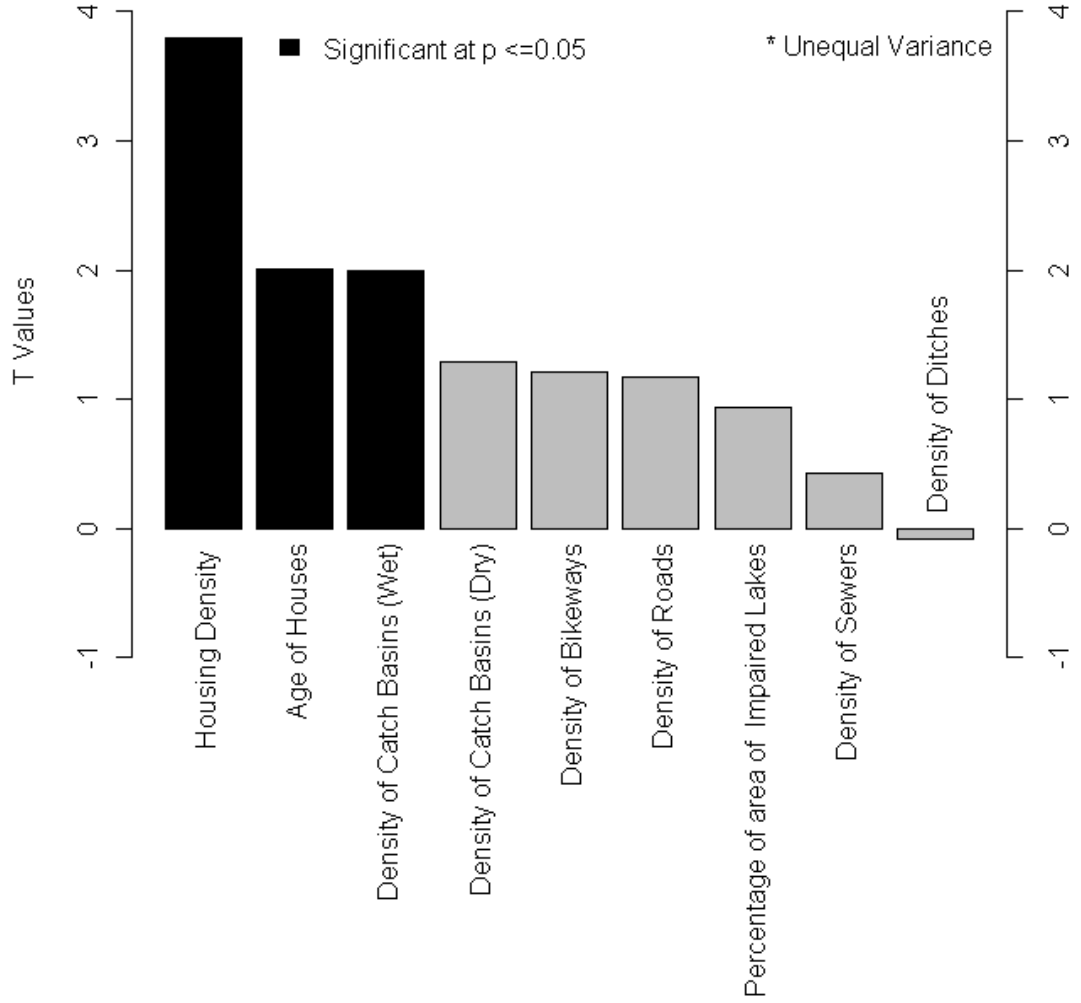
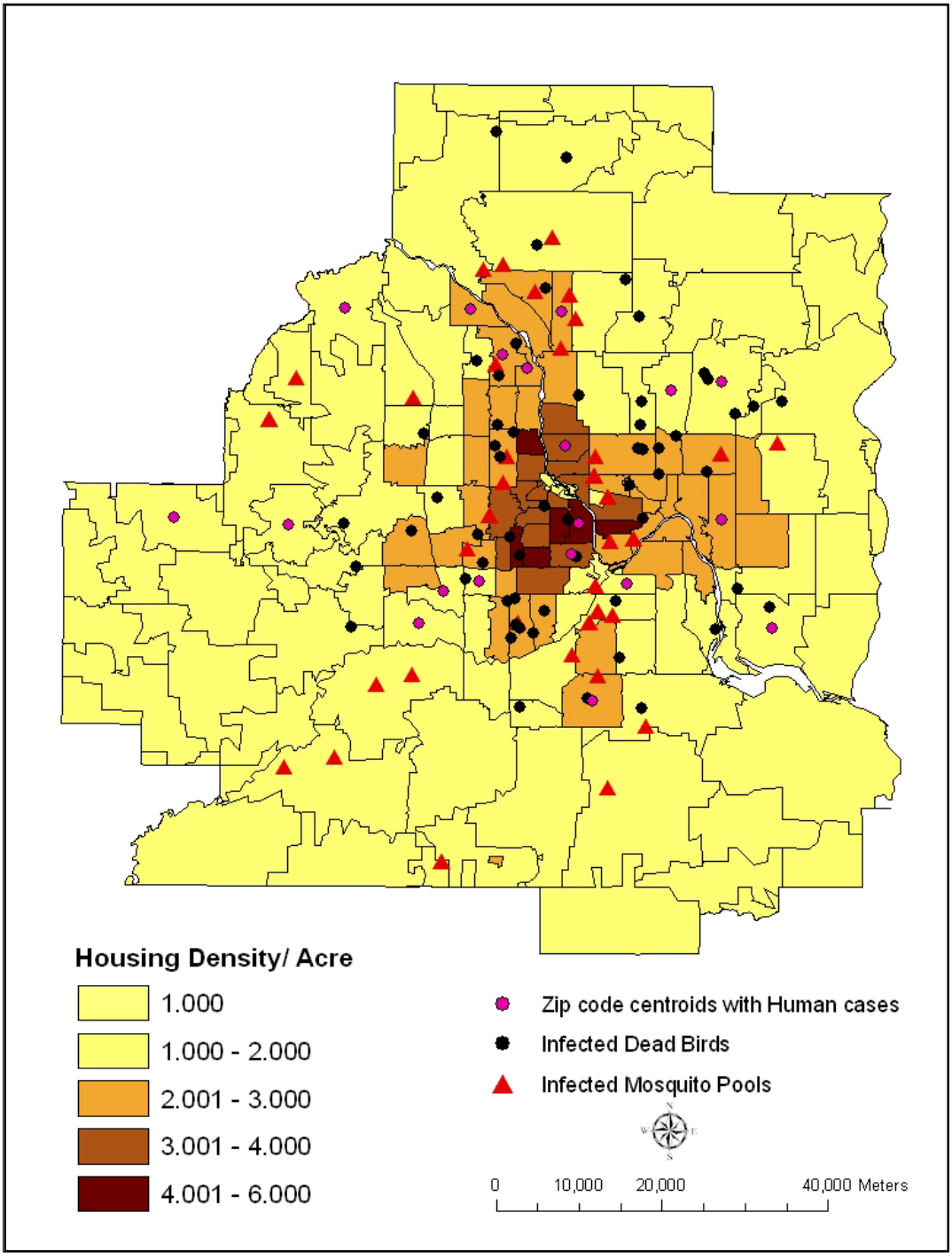


Figure 27 Spatial distribution of density of houses (per acre) in TCMA with West Nile virus illness in 2000



4.3.3 Proximity Factors

The t-values resulting from the t-tests comparing proximity to risk factors in zip codes with or without WNV disease reports are presented in Table 19 and Figure 28. Statistically significant lower mean values of all the proximity variables were found in zip codes with human incidences. The *distance to* variables in the order of descending t-values were distance to sewer lines (t-value = -3.9576), distance to golf courses (t-value = -3.5901), distance to open green space (t-value = -3.551), distance to bikeways (t-value = -3.4487), distance to water discharge sites (t-value = -2.9855), distance to lakes (t-value = -2.552), distance to impaired lakes (t-value = -2.5118), and distance to trails (t-value = -2.2244) (Figure 28). In addition, Figure 29 shows the negative spatial association between WNV disease concurrences in 2006, including infected dead birds, mosquitoes, and human cases with average distance to wastewater discharge sites.

Table 19 T-test results for Proximity Factors

Factors	t	p-value	95% Lower CI	95% Upper CI
Distance to Trails	**-2.2244**	**0.02755**	**-5520.51680**	**-327.62850**
Distance to Impaired Lakes	**-2.5118**	**0.01303**	**-2088.34130**	**-248.69300**
Distance to Lakes	**-2.552**	**0.0008081**	**-138.80250**	**-589.21500**
Distance to water discharge sites	**-2.9855**	**0.003285**	**-4843.41900**	**-986.44500**
Distance o Bikeways	**-3.4487**	**0.0007233**	**-2832.66640**	**-769.55380**
Distance to Open Green Space	**-3.551**	**0.0005068**	**-1595.10960**	**-454.84980**
Distance to Golf Courses	**-3.5901**	**0.0004416**	**-3177.02100**	**-921.87500**
Distance to Sewers Lines	**-3.9576**	**0.0001144**	**-6200.59100**	**-2071.89000**

Note: The mean of factors in **bold** are significantly different in zip code with human West Nile virus cases than zip codes without human cases at 95 percent significance level.

Figure 28 T-values for proximity variables based on Two Independent Sample t-tests

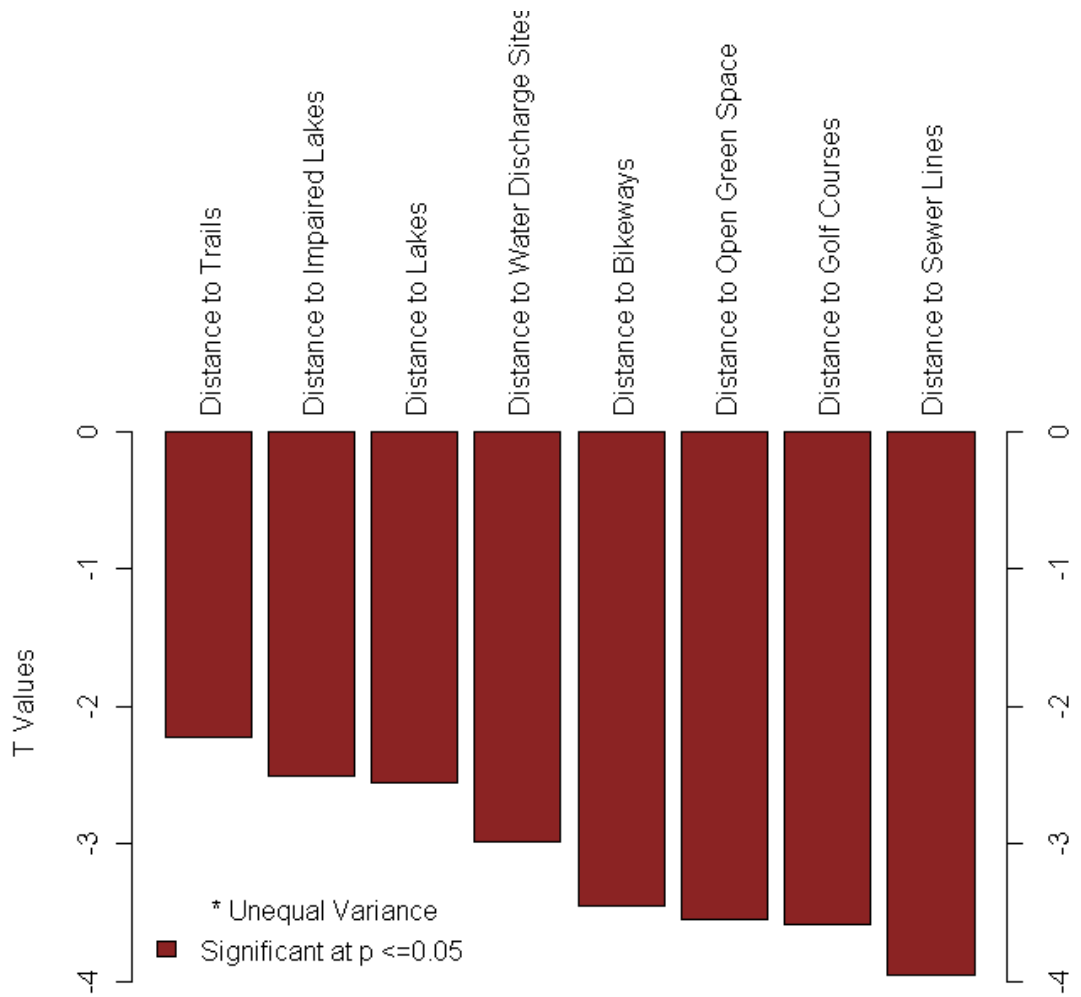
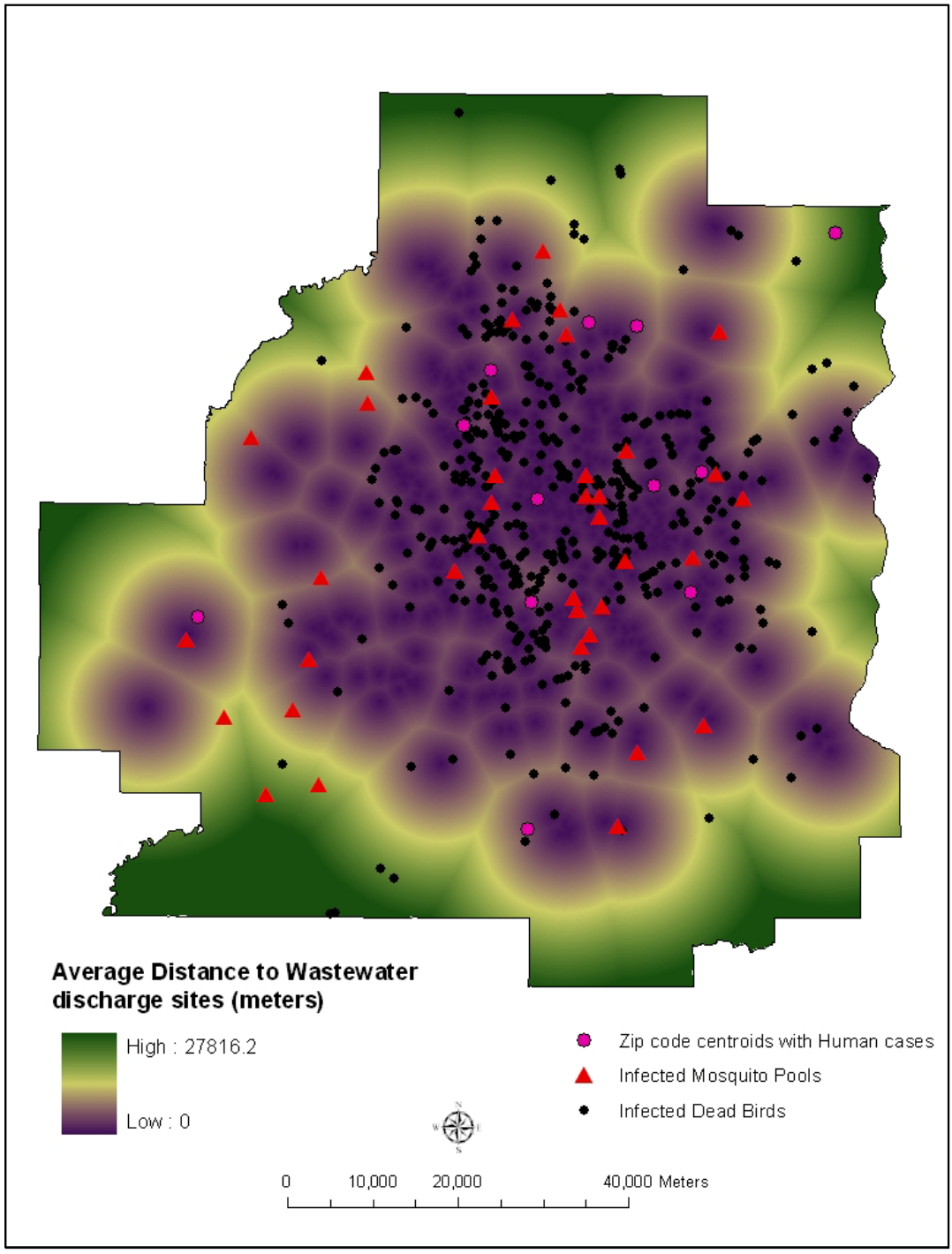


Figure 29 West Nile virus disease incidents in 2006 and average distance to wastewater discharge sites



4.3.4 Vector Control Policies

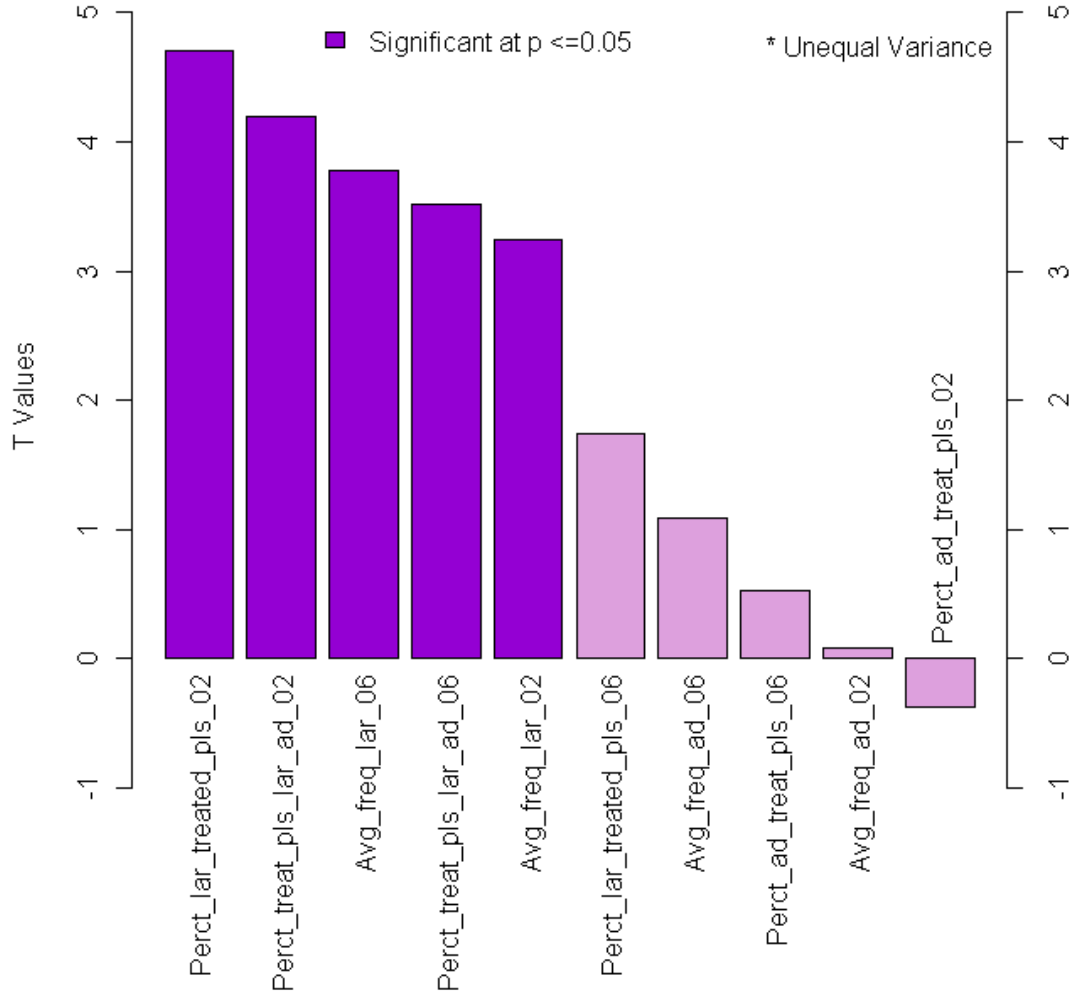
The test of means for vector control policy variables in zip codes with and without human WNV disease incidence is presented in Table 20 and Figure 30. For both the years of 2002 and 2006, the mean of frequency of larvicide treatment was statistically higher in zip codes with human infection. In 2002, the mean of frequency of larvicide treatment in zip codes with human cases was 31 times as compared to 16 times in zip codes without human cases. In 2006 the mean frequency remained similar (zip codes with WNV cases = 34 times and zip codes without WNV cases = 19 times). The mean of percentages of total treatment (both for larva and adult mosquitoes) were found significantly higher in zip codes with infected cases in 2002 and 2006. However, when analyzed separately, only the mean of percentage of larvicide treatment in 2002 was statistically different in zip codes with human cases (t-value = 4.7009). The t-tests of vector control policies showed interesting positive association with WNV disease incidents and will be discussed in the following section.

Table 20 T-test results for Vector Control Policies

Factors	t	p-value	95% Lower CI	95% Upper CI
Perct_treat_pls_lar_ad_02	**4.1985**	**0.0001598**	**14.56561**	**41.72092**
Perct_treat_pls_lar_ad_06	**3.5119**	**0.0005812**	**8.22220**	**29.35916**
Perct_lar_treated_pls_02	**4.7009**	**0.00000563**	**16.59050**	**40.63524**
Avg_freq_lar_02	**3.2379**	**0.002501**	**5.49517**	**23.82950**
Perct_ad_treat_pls_02	-0.3687	0.7128	-13.56746	9.29900
Avg_freq_ad_02	0.0855	0.932	-0.84467	0.92107
Perct_lar_treated_pls_06	1.7383	0.08413	-0.70392	11.03267
Avg_freq_lar_06	**3.7832**	**0.0005008**	**6.89509**	**22.69695**
Perct_ad_treat_pls_06	0.5265	0.5993	-6.51361	11.24828
Avg_freq_ad_06	1.091	0.2805	-0.3942256	1.33201

Note: The mean of factors in **bold** are significantly different in zip code with human West Nile virus cases than zip codes without human cases at 95 percent significance level.

Figure 30 T-values for Vector control policy variables based on Two Independent Sample t-tests



4.4 Discussion

West Nile virus infected dead birds, mosquito pools, and human cases have been reported throughout the TCMA, but with greater concentration in the urban/suburban areas of Minneapolis and Saint Paul. The mean values for most of the risk factors, which showed statistically significant difference between the zip codes with and without infected human case were directly and indirectly linked to the urban morphology of the area.

The association of the weather variables with WNV disease occurrence was positive. This finding was in line with the other studies, that is higher temperatures could potentially increase the probability of WNV occurrences (Hayes et al. 2005; Landesman et al. 2007; Zou et al. 2007). However this was contradictory to the study based at Iowa where the mean annual temperature was consistently lower in census block groups with human cases from the year 2003 to 2006 (DeGroote et al. 2008). This negative association between temperature and WNV illness could be a result of using coarse annual weather data. On the contrary, detailed weather data at the diurnal resolution were used for this study. This effectively reflected the hypothesized positive association between temperature and the occurence of WNV cases.

The t-test results of the land cover variables supported the urban/suburban nature of disease incidence in TCMA. Among the 14 land cover classes, statistically significant higher mean values of developed open space, low density, medium density, and high density land cover classes were found in zip codes with WNV incidents in birds, mosquitoes, and humans. In addition, the t-values showed significant negative association between pasture/hay and cultivated crops and disease cases. This further contributed to the urban-centric nature of WNV disease transmission in the TCMA. These research findings from the land cover variables were also supported by some of the built-environment variables. Higher and significant mean values of housing density, age of houses, and density of urban catch basins (wet) in zip codes with WNV human cases than zip codes with no case also pointed to the urban preference of WNV occurrence in the TCMA. The *Culex restuans, and Culex pipens* vector species are two

of the four important drivers of WNV transmission in TCMA. They are predominantly urban mosquitoes and found in the populated residential areas. Given the urban preference of WNV transmission in the TCMA, these vector species can play important roles in the transmission of the virus from mosquitoes to humans. A study focused on mosquito surveillance and WNV in Connecticut indicated that *Culex restuans* appeared to be important in initiating the virus transmission among birds in early summer and *Culex pipiens* played a greater role in amplifying the virus later in the season (Andreadis et al. 2001). The urban/suburban nature of WNV occurrence in TCMA is similar to the WNV outbreak in Chicago in 2002 (Ruiz et al. 2004; Ruiz et al. 2007), Detroit metropolitan area in 2007 (Ruiz et al. 2007), and in Georgia (Gibbs et al. 2006). In Chicago and Detroit, areas classified as 'Inner Suburbs' had the highest percentage of WNV disease cases. Similarly in Georgia, Gibbs *et al.* found that the rate of WNV infection increased in urban/suburban areas with high housing density and decreased in the mountainous region of the state (Gibbs et al. 2006).

On the other hand, the association of WNV illness with urban settings of the TCMA contradicts other group of studies, which show a positive association of rural landscape with WNV human cases (Reisen et al. 2004; Miramontes et al. 2006; DeGroote et al. 2008). A recent study concluded that rural agricultural setting, including irrigated areas and animal feeding operations were strongly associated with human WNV disease incidences (DeGroote et al. 2008). Another study based in California suggested that farmhouses created 'islands' or pockets of elevated vegetation, attracting both birds and *Culex tarsalis* mosquito species for nesting and host-seeking activities respectively (Reisen et al. 2004). This increased the risk of WNV transmission. Agricultural crop sales were a significant predictor of WNV incidents at a county level study in Colorado, Nebraska, Louisiana, and Pennsylvania from 2002 to 2003 (Miramontes et al. 2006).

Among the built environment variables, housing characteristics and density of urban catch basins emerged with higher statistically significant mean values in zip codes with infected human cases. Unlike any other metropolis (Chicago or Detroit) in

the Mid Western United States, the TCMA is home to many lakes, which have become an important feature of the urban landscape. To protect the lakes from polluted urban run-off water, storm water drains and catch basins were constructed to store run-off water. These catch basins, often with stagnant water and decaying plant residues, could likely be potential breeding grounds for *Culex* mosquitoes and therefore increase the risk of WNV infection. This is especially true in times of drought when the water stands for a longer period of time (Shaman et al. 2005). A positive association between housing age (50 to 60 years old) and WNV occurrence was an interesting result and warranted further investigation. Housing age also emerged as a dominant factor in the discriminant analysis of the Chicago study and was strongly associated with WNV outbreak in 2002 (Ruiz et al. 2004).

This study made a significant contribution to the literature of WNV by investigating the association between disease occurrence and *distance to* hypothesized risk factors. Other studies had measured the risk factors as percentages or presence and absence in an area. This study went a step further and attempted to analyze an important question: how does proximity or spatial distribution of a risk factor influence the risk of WNV. As discussed in section 4.3.3, statistically significant lower mean values of proximity variables were found in zip codes with WNV infected human cases (Table 20). This implied that there was a negative association between disease occurrence and distances to lake, open green space, trails, impaired lakes, water discharge sites (Figure 29), bikeways, golf courses, and sewer lines.

Since *Culex* species can transmit WNV as well as Western Equine Encephalitis (WEE), surveillance for these species was revamped with the emergence of WNV in 2002. MMCD, the principal vector surveillance agency in TCMA, traps larva and adult mosquitoes for testing by four methods: sweep net collection, CO_2 trap collection, New Jersey light traps, and gravid traps. MMCD uses sweep net collections to monitor human annoyance during the peak mosquito activity period. CO_2 traps baited with dry ice are used to monitor mosquito population levels and the presence of disease. New Jersey light traps are used to compare mosquito species population levels from year to

year. In addition to CO_2 traps, gravid traps are also used to monitor *Culex* adults. The gravid trap is designed to attract female mosquitoes that are seeking oviposition sites. There are two mosquito control programs for *Culex* species, larviciding and adulticiding. Larva control is the main focus of the program but is supplemented by adult mosquito control when required. For both the programs, vector control activities are recorded and reported by Public Land Survey (PLS) units. Based on t-test results, higher and significant mean values of percentages of PLS units treated for larvicide and adulticide in 2002 and 2006, percentages of PLS units treated for only larvicide (2002), average frequency a PLS treated for larva control (2002), and average frequency a PLS treated for adult control (2006) were found in zip codes with infected human cases. This indicated a positive association. Spatially, the vector control programs are concentrated more in the populated areas of TCMA than the surrounding rural areas. Typically we would expect that with more efficient vector control programs there will be lesser number of WNV cases, i.e. a negative association. However the positive association of vector control programs in TCMA was due to the focus and priorities set by the MMCD. MMCD uses “Priority Zones” to focus service in areas where it will benefit the highest number of citizens. Priority Zone 1 contains the majority of the population of the Twin Cities metro area and has boundaries similar to the Metropolitan Urban Service Area (MUSA). Priority Zone 2 includes sparsely populated and rural parts of the TCMA. Small towns or population centers in Priority Zone 2 are considered satellite communities and receive services similar to Priority Zone 1 (MMCD 2004). Hence we see a positive association between vector control variables and WNV disease occurrence, especially in the populated parts of TCMA.

4.5 Conclusion

The dynamics of WNV illness in birds, mosquitoes, and humans showed strong association with urban landscape features in the TCMA. Some of the urban characteristics associated with disease incidences were developed land cover classes (developed open space, developed high, medium, and low density), higher housing density, houses built 50 to 60 years before, higher density of urban catch basins and storm water ponds, and smaller distances to pockets of natural areas within the city, such as lakes, bogs, open green space, swamps, and trails. These findings suggest that *Culex restuans* and *Culex pipiens*, which are mainly urban mosquitoes, might play major roles in the transmission and amplification of the virus.

With this *exploratory* analysis of the association of potential risk factors with WNV illness, I then proceed to the next chapter where I attempt to tease out the spatial association between *only* urban landscape features and WNV disease incidences *rigorously*.

5. Chapter 5: How urban landscape features in the Twin cities metropolitan area of Minnesota associate with the transmission of West Nile virus?

5.1 Background

West Nile virus infection in birds, mosquitoes, and humans has exhibited strong spatial clustering in urban areas of the Midwestern United States. For example, in 2002, the states like Illinois and Michigan led the nation with 884 and 664 infected human cases respectively. Spatially, most of these disease incidents were clustered around the urban areas of Chicago and Detroit. Similarly In 2003, when the WNV infection reached a level of epidemic in Minnesota (148 human cases and 433 infected dead birds) (MDH 2003), significant spatial clusters of infected dead birds and human cases were found in the urban areas of Minneapolis and Saint Paul.

The WNV outbreaks in the urban landscape occur in a diverse mixture of buildings, sprawling development, transportation routes, vegetation (open green space, parks, trees, shrubs etc.), pockets of natural areas (lakes, reservoirs, golf courses, unpaved trails, etc), and people. These landscape features are associated with myriad aspects of urbanization which in turn affects urban ecosystem health, conservation, and human-social phenomena like economic activities, crime patterns, and *human health*. However the role of these contextual factors is often neglected in research related to disease transmission. It is important to understand how these urban features affect

amplification, transmission, and disease pattern. This is especially critical for multi-host pathogens such as WNV, where birds and mosquitoes and their interaction with the natural and built environment in urban areas are fundamental to disease transmission.

Recently Ruiz *et.al* conducted an comparative analysis of association of WNV infected human cases and urban landscape features in Chicago and Detroit (Ruiz et al. 2007). The authors divided the two metropolitan areas into five urban classes, with Central Business District (CBD) at the center and exurbs in the outlying areas. In both the metropolitan areas, WNV cases were significantly higher in the Inner Suburb class. The dominating urban features of this class included high housing density, 1940 – 1960 period housing, and moderate vegetation cover (Ruiz et al. 2007). Given the similar nature of WNV transmission in the TCMA, the focus of this chapter is to extend Ruiz *et al*'s study. The specific objective is to investigate and analyze the association of specific urban landscape features that contributed to the viral activities of WNV infection in the TCMA from 2002 - 2006.

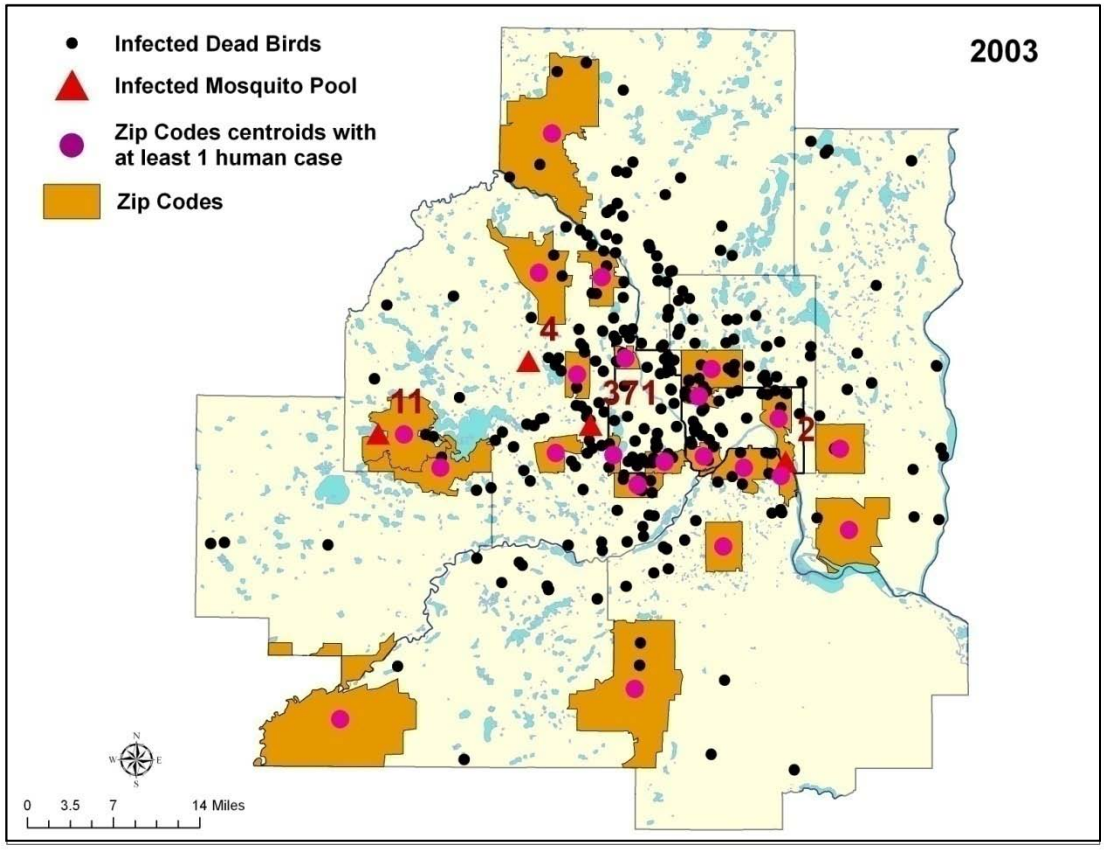
There were several reasons to follow Ruiz *et al's* study in the TCMA. First, during the WNV epidemic year of 2003, the number of confirmed cases in the TCMA were 285 infected dead birds, 1400 infected mosquito pools (approximately), and 26 human cases. Even though the numbers were low compared to cases in Chicago and Detroit, the spatial distribution of incidences showed strong spatial clustering in the urban areas of TCMA (Figure 31). Second, the ongoing urbanization in the TCMA has profound implications to shape the environmental and socioeconomic characteristics of the region, which in turn could influence human health. This 7,700 km^2 seven-county area is the economic hub of a multistate region. Home to 2.8 million people, and forecasted to top 3.5 million by 2020, it is also a major center of sprawl (Ghosh and Manson 2008). The rapid expansion of rural areas into urban, suburban, and exurb agglomerations, buffered from others by undeveloped land is constantly changing the urban morphology of the TCMA. Third, the landscape of TCMA is an ideal setting for examining the association of urban features and disease transmission. It stands in relative isolation from other large urban agglomerations, making it easier to extract the

relationship between urban heterogeneous environment and human health at a metropolitan scale.

Fourth, similar spatiotemporal trend of WNV occurrences were found in Chicago, Detroit, and TCMA. For example an epidemic year in Chicago and Detroit (2002) was preceded by very low number of WNV cases in the previous year (2001). This was also the trend in TCMA, where the number of infected dead birds and human cases doubled from 2002 to 2003 (epidemic year). Further the number of cases declined in 2004 and 2005 for all the three urban regions and showed an increasing trend from 2006. Fifth, the urban areas in the upper Midwestern United States have similar climatological and geologic features. The glacial drift and massive retreat of ice sheets about 75,000 years ago resulted in glacial deposits, erosion, lakes, and rivers. The current physiography and soil characteristics in the TCMA were created by the most recent drift of Wisconsin glaciations. The TCMA falls under the "drift area" with relatively flat terrain and moderate to poor drainage systems. For example, in Anoka County, the Anoka Sand Plain that covers most of the area has numerous depressions and old glacial drainage ways. At present these depressions are lakes or bogs (USDA 1977). The other parts of TCMA are also flat terrain with sandier soils, making these areas prone to flooding and standing water similar to that of Chicago and Detroit.

This chapter is organized as follows. The section 5.1.1 briefly sets the background of urban growth models, factorial ecology, and effects of urban landscape on different health outcomes. The section 5.2 describes the data and the methodology used, section 5.3 reports the results, section 5.4 analyses the results, and finally section 5.5 concludes the chapter.

Figure 31 Spatial distribution of WNV illness in the Twin Cities Metropolitan Area of Minnesota in 2003



5.1.1 Urban Morphology and Health

Geographers and sociologists have analyzed urban morphology since the early twentieth century. The studies mainly focused on location of CBD and resulting urban growth from this focal point (Kaplan et al. 2004; Hall 2006). Around the 1920s three traditional urban growth models emerged from the sociology department of University of Chicago. First, the Concentric Zone model or Burgess Model hypothesized that the city expanded in concentric rings from the CBD. Second, the sector or Hoyt's model

modified the Burgess model such that urban form did not develop in concentric zones but along transportation routes into sectors of differential sizes from the CBD. Third, Multiple Nuclei model described the pattern of a city as fragmented urban spaces centering around multiple nuclei serving distinct functions (Lappo et al. 1992; Kaplan et al. 2004; Hall 2006).

Later, around 1960 to 1970, *urban factorial ecology* approach was used in analyzing the characteristics and patterns of city growth. A geographer, Brian Berry employed factorial ecology approach in several studies involving urbanization, patterns of city growth, and urban land use (Berry 1971). In this approach, factor analysis reduces and organizes a large number of data related to physical, social, economic, and demographic characteristic of a city to meaningful factors and indices, which were then used in regional classification schemes. The technique could be applied for both exploratory analysis and hypotheses testing. In recent years with the availability of cheaper and sophisticated computers, efficient multivariate statistical analyses, and volumes of data, factorial ecology approach is frequently used in studies to develop social metrics and indices to better understand economic activities, land use change (Herold et al. 2005), landscape characterization (Owen et al. 2006), and ecosystem functions and health (Alberti et al. 2003; Alberti 2005).

Research on relationship between urban form and health is not new. In one of the earliest studies on environment and health, Faris and Dunham showed that psychiatric admissions in Chicago varied with location of residence within the city. Higher rates were observed among those living in the inner-city core than in outlying areas (Faris and Dunham 1939). More recently there had been several studies linking urban sprawl, congestion, traffic, high density of buildings, urban waste, disturbance of natural areas within a city, and pollution (air, water, and noise) to various health effects such as mental health (Leventhal and Brooks-Gunn 2003; Gary et al. 2007; Clark et al. 2008; Howell and McFeeters 2008; Watson et al. 2008), obesity (Ewing, Schmid et al. 2003; Saelens et al. 2003), water related infections (Greenberg et al. 2003; Aramaki et al. 2006; Drechsel et al. 2008), pedestrian-vehicle accidents (Ewing, Schieber et al.

2003), asthma, and other respiratory ailments (Oudinet et al. 2006; Kyrkilis et al. 2007; Grineski 2008; Jones et al. 2008; Schikowski et al. 2008).

Transmission of infection and diseases by multi-host pathogens, namely malaria, Lyme disease, or WNV are also associated with patterns of vegetation, hydrology, and human settlements in an urban area (Kitron 1998; Smith et al. 2004). For example, there are several studies that provide evidence at various spatial scales that malaria and encephalitis-producing viruses are associated with host of environmental, built environment, social, and demographic factors that affects their habitat characteristics (Rogers and Randolph 2000; Barrera et al. 2001; Sattler et al. 2005). Similarly, it is important to understand the association between urban form and vector-borne disease such as WNV illness, especially where birds and mosquitoes and their interaction with the natural and built environment in urban areas are fundamental to disease transmission.

5.2 Data and Methodology

5.2.1 Data

The hypothesized risk factors of WNV transmission in the TCMA area were divided into four categories of environmental, built-environment, proximity, and mosquito abatement policies (refer Chapter 2 and Chapter 4). Among these, factors which described or were part of the Twin Cities' urban landscape and morphology were included in this analysis. The selected variables are listed in Table 21. The incidence data of WNV infected dead birds, mosquito, and human cases from 2002 to 2007 were also considered. Both the incidence and urban risk factor data were aggregated at the zip code level (refer Chapter 2 for details on data processing).

Table 21 Description of urban landscape features hypothesized to be associated with West Nile virus transmission

Categories	Factors
Environment	Land Cover (14 classes), Density of Streams/sq. mile, Elevation
Built-Environment	Density of urban catch basins/sq. mile (dry), Density of urban catch basins/sq. mile (wet), Density of Ditches/sq. mile, Housing density/acre, Age of houses, Density of Roads/sq. mile, Density of Population
Proximity	Distance to 8 types of Wetlands, Distance to lakes, Distance to Open Green Space, Distance to sewers, Distance to waste water discharge points, Distance to Streams, Distance to Golf Courses, Distance to Trails, Distance to Bike paths, Distance to Impaired lakes

5.2.2 Methodology

The urban factorial ecology approach, explained briefly in section 5.1.1 was employed here. This approach uses Principal Component Analysis (PCA) to derive uncorrelated components from a set of correlated variables. These components quantify myriad characteristics of urban neighborhoods ranging from socioeconomic to built environment features (Berry 1971).

PCA is a powerful multivariate statistical technique that can be used to simplify a dataset by reducing the number of correlated variables into a smaller number of uncorrelated principal components (PCs). It transforms or rotates the entire dataset to a new coordinate system such that the greatest variance by any projection of the dataset is described by the first axis or the first principal component. The second greatest variance is then explained by the second component and so on. The advantages of using PCA are

as follows: 1) it reduces the dimensionality of a dataset by retaining PCs which explains the maximum amount of variation, 2) because the PCs are uncorrelated, multicollinearity can be avoided by using the components in place of the original variables, and 3) it is an exploratory tool to identify patterns and relationships among groups of related variables. In spite of such usefulness, PCA has disadvantages. Its greatest disadvantage lies in the subjective interpretation of the PCs. In situations where different (theoretically) variables are highly correlated and contribute significantly to one component, describing such a component becomes difficult. In other words, it is not possible to know what the 'components' actually represent. In such a situation only theory and domain knowledge of the research problem can help to inform the researcher on this.

In this analysis three groups of variables were included that described environment and built-environment features conducive for the transmission of WNV in the TCMA (Table 21). The environmental factors were: land cover categories (14 classes), density of streams/sq. mile, and elevation. The variables in the built environment category were density of urban catch basins/sq. mile (dry), density of urban catch basins/sq. mile (wet), density of ditches/sq. mile, housing density/acre, age of houses in years, density of roads/sq. mile, and density of population. Lastly, the proximity variables included distance to 8 types of wetlands, distance to lakes, distance to open green space, distance to sewers, distance to waste water discharge points, distance to streams, distance to golf courses, distance to trails, distance to bike paths, and distance to impaired lakes.

The dimension of this dataset with 40 correlated variables were reduced to a much smaller number of uncorrelated components using the *Principal Component* function in the ‘stats’ package of R statistical programming language. A varimax rotation was used to rotate the components. The next important step in the methodology was to extract the relevant PCs for further analysis. The goal here was to retain the number of components that accounted for as much variation as possible with the fewest meaningful components. Typically a subset of $k < p$ of components are selected by a

three-part process (p = total number of PCs and k = selected PCs). First, the number of PCs is retained by examining the slope of the 'scree-plot'. A scree plot shows the eigenvalues on the y-axis and PCs on the x-axis. The eigenvalues describe the amount of variance explained by each component. Second, usually PCs with eigenvalues greater than one are selected. Third, through sequential selection, PCs that explain 90 percent of the cumulative variation of the entire data set are retained.

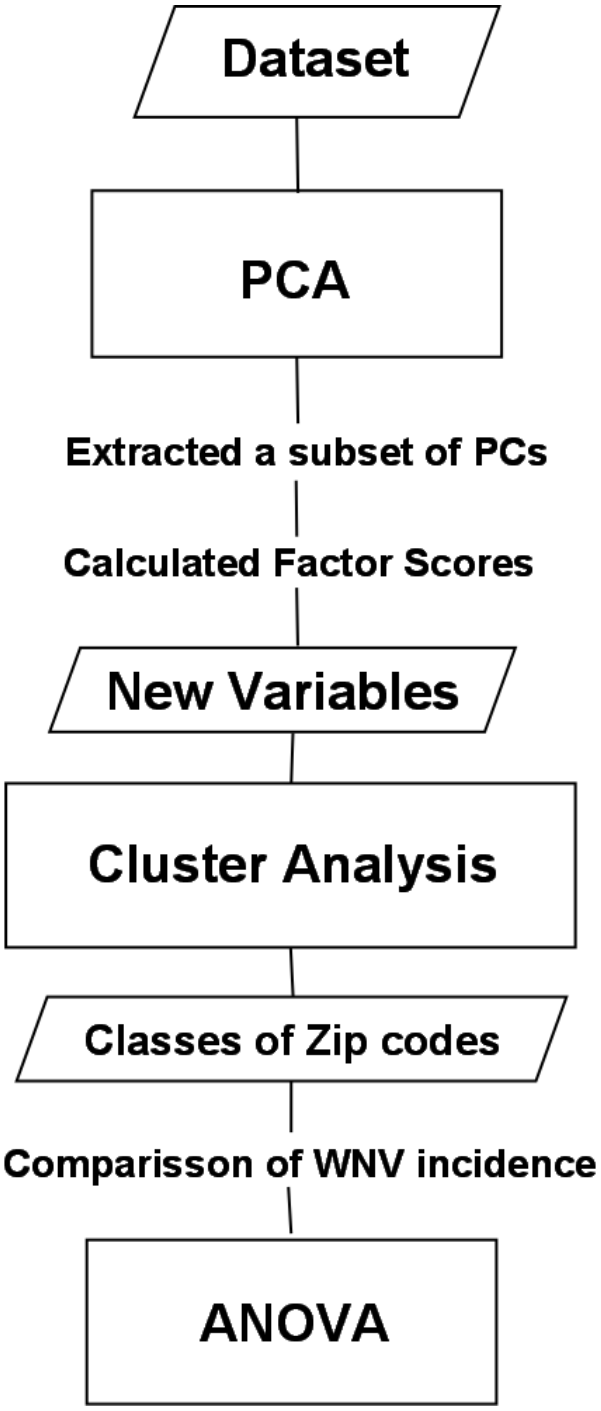
Finally, using the selected PCs, factor scores were calculated for each record (zip codes) for further analysis. There are a number of different methods that can be used for estimating factor scores from the PCA output. These include: Ordinary least squares, Weighted least squares, and the Regression method. For this analysis, the regression method was used, which involves calculating the maximum likelihood of factor loadings of the original variables for the retained PCs. For each of the selected PCs, the factor loadings express the correlation of original variables with that particular PC. Therefore in a regression method, the factor score for each observation is estimated by supplementing the observed data by a vector of factor loadings for the i^{th} observation.

In the next step, hierarchical agglomerative cluster analysis was conducted with the factor scores as input data. The output of the cluster analysis was classes with zip codes showing similar urban landscape characteristics. I used the *hclust* function from the 'stats' package of R statistical programming software to perform cluster analysis. This function performs a hierarchical cluster analysis using a set of dissimilarities for *n* objects being clustered in two parts (Anderberg 1973; Gordon 1999). First, each object (say zip code) is assigned to its own cluster or node in a tree, and then the algorithm proceeds iteratively, at each stage joining the two most similar clusters or nodes. This continues until there is just a single cluster. A number of different clustering methods are provided. For this particular clustering algorithm, *Ward's* minimum variance method is used, which aims at finding compact and spherical clusters. Second, the *cutree* function was used to group the clusters or nodes of the tree into smaller number of clearly defined classes. These derived classes represented the urban landscape in the

TCMA.

For the next part, analyzing the association between derived urban classes and WNV infected mosquitoes and human cases; I computed two incidence rates per 100,000 people. First, incidence rate of mosquitoes in the WNV infected pools and second, incidence rate of human cases. An exploratory spatial data analysis (ESDA) approach, *overlay*, was conducted to explore the spatial association of WNV infected mosquito and human incidence rates with the urban classes. Further, in order to assess the differences in the degree to which WNV illness affected the different urban classes, ANOVA was computed on the zip code means of the incidence rates for the different classes. The null hypothesis was H0: There was no difference of WNV incidence rates among the urban classes. Since the Levene's test for homogeneity of variance revealed non-constant variances, Brown-Forsythe ANOVA was used instead. This function does not assume normal distribution and homogeneity of variance. Figure 32 highlights the important steps of the methodological framework described above.

Figure 32 Flow diagram showing the methodological framework



5.3 Results

A three step process was used to classify the TCMA into homogenous urban classes. First, PCs were selected explaining maximum variation of the dataset. Second, new variables were calculated factor scores from the selected PCs. Third, the new variables became the input variables for the hierarchical cluster analysis, which divided the TCMA into five urban homogenous classes.

The PCA with varimax rotation created 40 uncorrelated PCs from the original 40 risk factor variables. Table 22 shows the eigenvalues and cumulative variation explained by the first 10 PCs. Of these, the first five PCs together accounted for 84 percent of the total variation. The first PC explained 40 percent and the second 21 percent. There was a significant drop in the amount of variance explained from PC1 to PC2. Also, beyond PC6, the eigenvalues of the remaining components dropped below value one and their respective variances also declined below one percent.

Table 22 Variance Explained by the selected Principal Components

Principal Components	Eigen Values	Variation (%)	Cumulative Variation (%)
1	**12.000**	**40.40**	**40.40**
2	**7.240**	**21.10**	**61.500**
3	**2.620**	**10.00**	**71.500**
4	**2.070**	**7.30**	**78.800**
5	**1.700**	**5.20**	**84.000**
6	0.470	1.10	85.100
7	0.360	0.90	86.000
8	0.180	0.06	86.060
9	0.120	0.04	86.103
10	0.060	0.01	86.113

Note: The variances for the selected principal components are in **bold**

Figure 33 Scree plot showing the Eigen Values obtained from the Principal Component Analysis with 40 urban variables

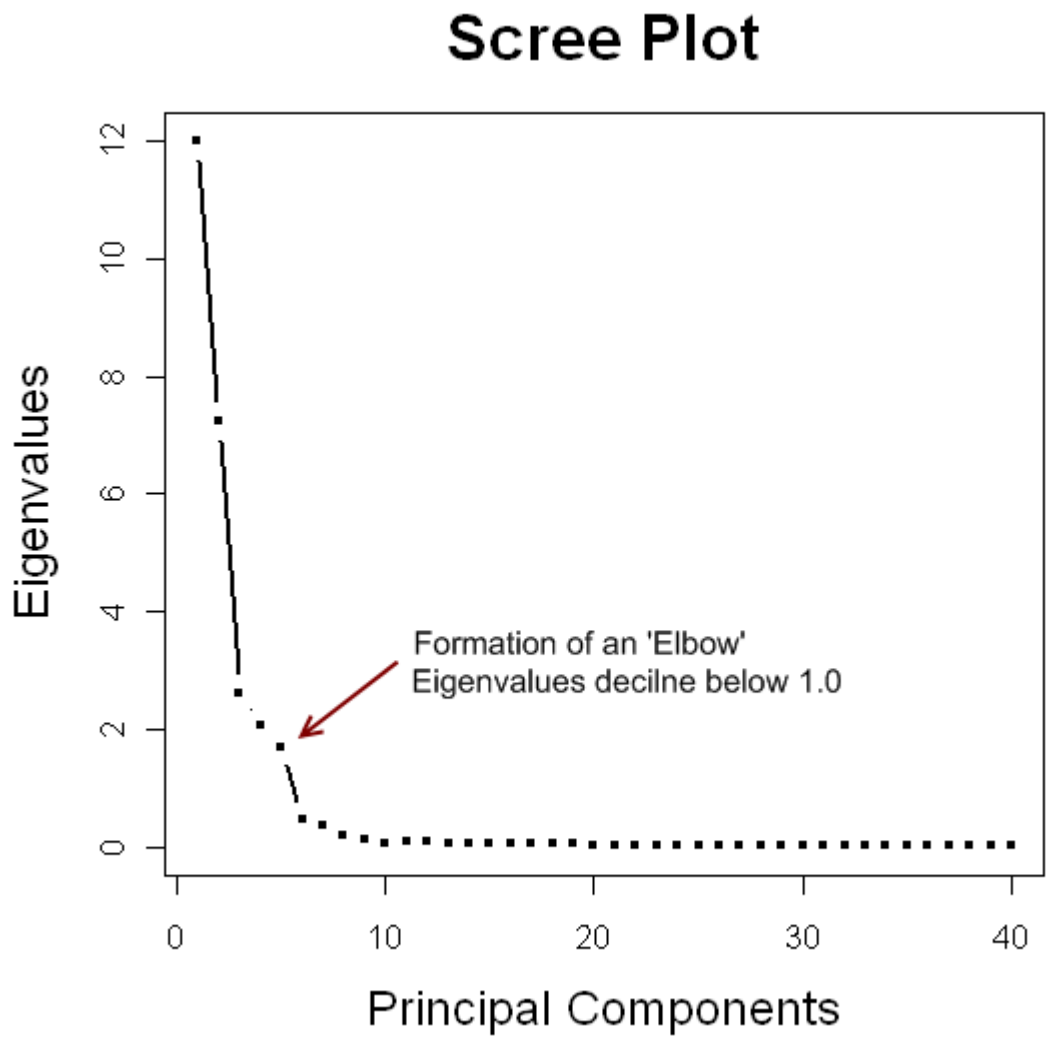


Figure 33, which is a scree plot, also confirms the drop in the eigenvalues beyond PC5. When read left to right across the X-axis, this plot showed a clear separation between PCs with high-explained variance versus low-explained variance at

PC5. The point of separation is termed as *elbow*. Thus, based on Table 22 and Figure 33, the first five PCs were retained for further analysis. They cumulatively explained 84 percent of the variation with PC1 explaining a large majority, followed by PC2, PC3, and so on. The formation of an 'elbow' at PC5 indicated the likely separation of the most important PCs from the less important PCs.

The variables contributing significantly (higher positive and negative factor loadings) to each of the selected components are as follows:

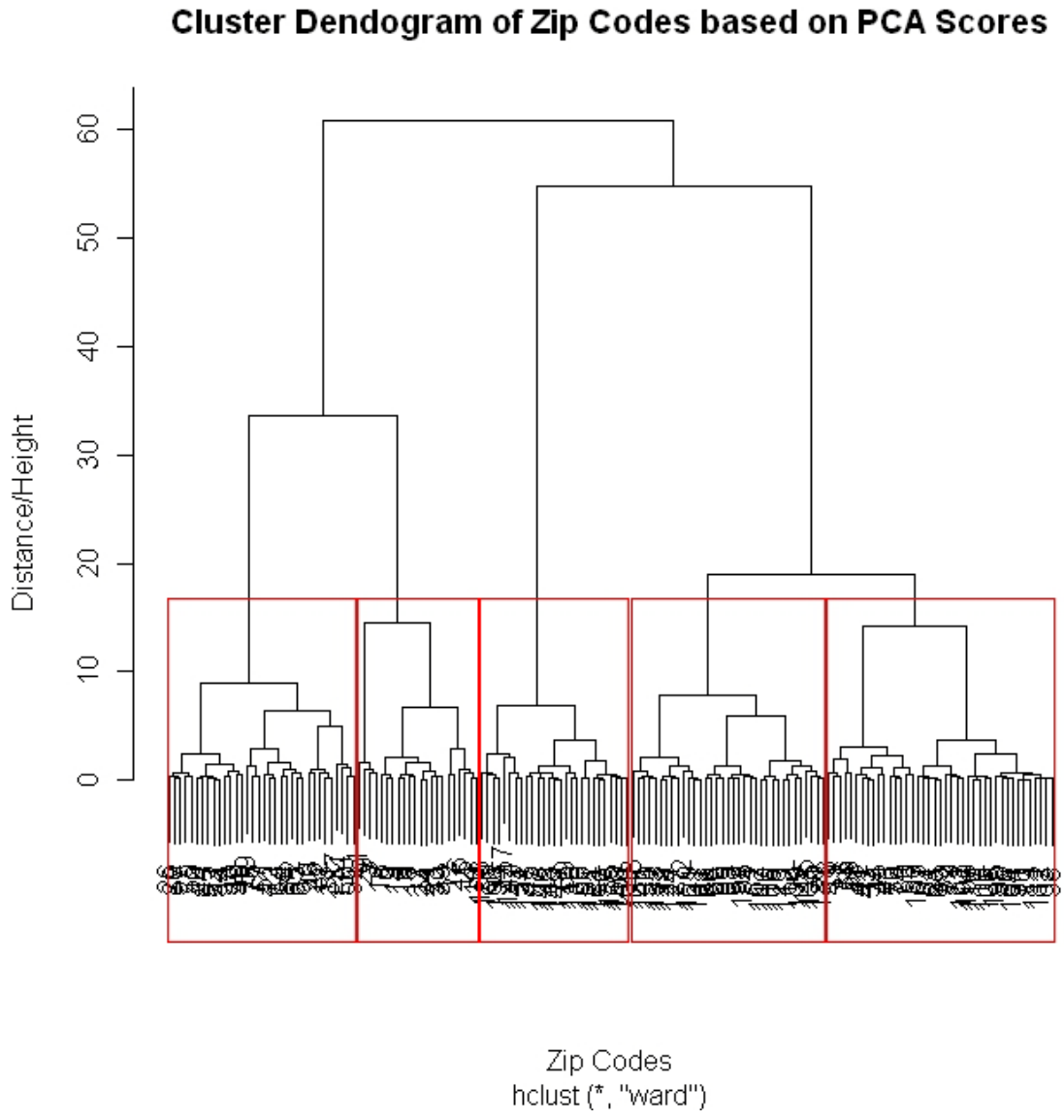
1. **PC1** – developed-low density land cover class, developed-medium density land cover class, high housing density, housing from 1940s-60s, population density, high density of catch basins, and smaller distance to parks, sewers, trails, waste-water discharge points.
2. **PC2** – developed-high density land cover class with impervious surface accounting for 80-90 percent of the total cover, larger distances to natural areas including lakes, wetlands, and open green space, usually occupied by commercial/industrial buildings, and lower density of single family residential houses.
3. **PC3** – medium population density, and recent development of houses (1980s – 90s) with absence of older houses
4. **PC4** – low population density, high diversity of land cover classes, and vegetation
5. **PC5** – Land cover classes namely cultivated croplands and pastures/hay.

In the next step, factor scores were calculated for each PC using the associated factor loadings of the variables mentioned above. The output was five new variables: Score1, Score2, Score3, Score4, and Score5, which represented the features of 40 original variables.

Finally, cluster analysis was conducted on these new variables to obtain five urban landscape classes. Figure 34 is a dendogram, which is a graphical representation of hierarchical clustering based on similarities and dissimilarities of factor scores obtained from the PCA. The red rectangles showed the grouping of zip codes into five

groups. Here parsimony was achieved with a small number of clearly defined urban classes.

Figure 34 Dendogram showing the hierarchical clustering of zip codes into 5 classes of urban landscape in Twin Cities Metropolitan Area



The five urban landscape classes and their dominant characteristics are as follows:

1. **City, High Density** – high density of urban catch basins, open green space within a distance of 1 mile, high housing density, some 70 - 75 years old housing, smaller distance to sewers, and developed high density of land cover class.
2. **City, Medium Density** – presence of wetlands like swamps and bogs, high density of urban catch basins, open green space within a distance of 0.5 mile, housing belonging to 1940-60s period, and developed medium density land cover class.
3. **Suburb** – developed low density land cover class, low housing density, presence of more natural features like lakes, parks, shallow fresh marsh, and diverse vegetation.
4. **Outer Suburb 1** – Recent development houses between 20 – 25 years old, shallow fresh marsh, lakes, and diverse vegetation in the form of shrubs, pastures, and deciduous forest.
5. **Outer Suburb 2** – Agricultural area, pasture, open green space, and deciduous forest land cover.

To determine the degree to which the urban classes defined above were in tune with the characteristics of original variables, I further calculated the mean values of the dominant variables for each of the five urban landscape classes. The results are summarized in Table 23.

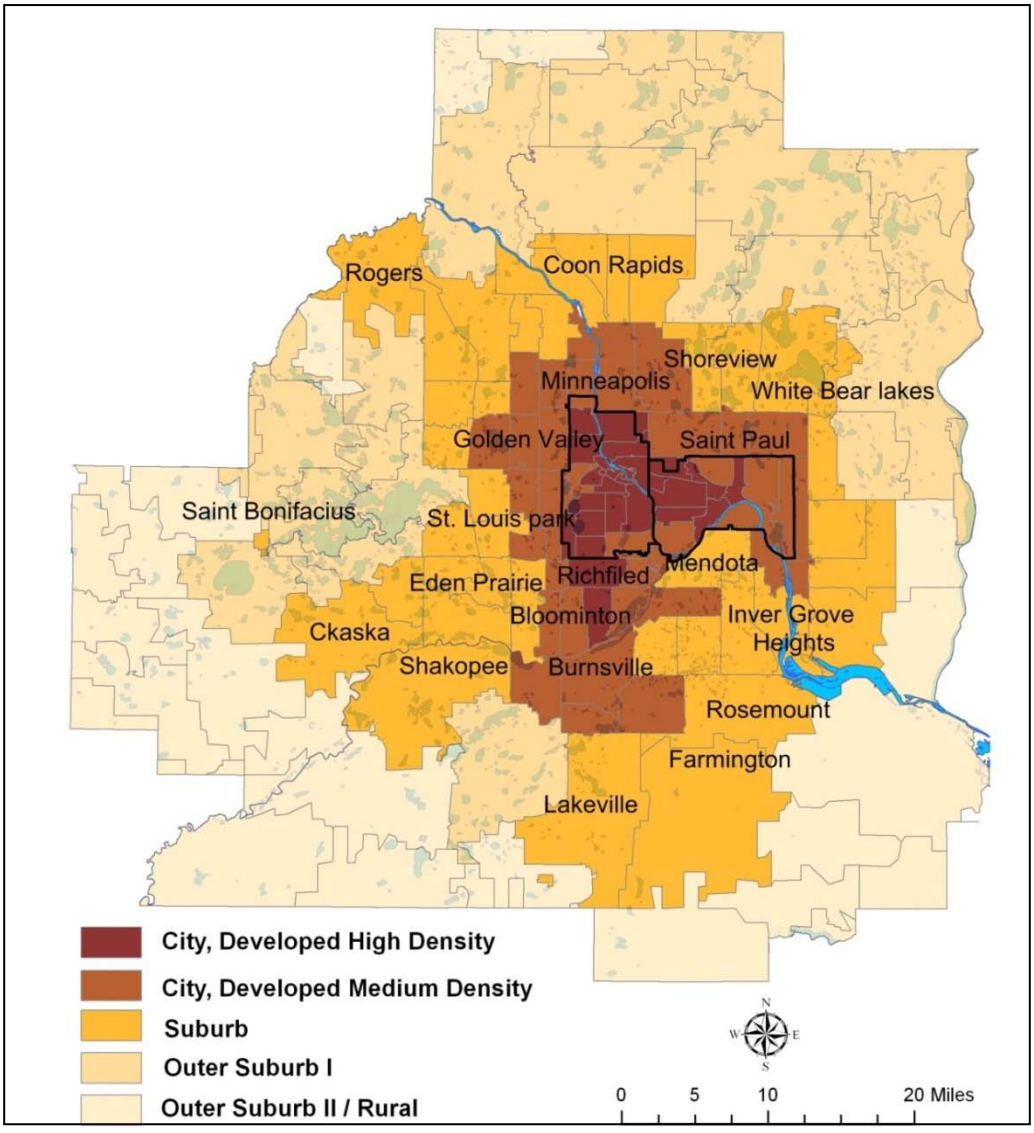
Table 23 Mean values of the original variables which emerged dominant in defining the urban classes

Variables	City-High Density	City-Medium Density	Suburb	Outer Suburb 1	Outer Suburb 2	Unit
Distance to Shallow Marsh	1.03	0.40	0.31	0.31	0.61	mile
Distance to Shrub Swamp	5.36	3.27	2.29	1.54	3.81	mile
Distance to Wooded Swamp	6.93	4.92	6.03	3.27	6.99	mile
Distance to Bog	3.69	2.80	3.73	4.09	9.30	mile
Density of Catch Basins (Wet)	102.51	66.15	34.49	7.86	1.38	sq/mile
Density of Catch Basins (Dry)	514.13	271.44	159.50	26.56	6.24	sq/mile
Distance to Lakes	0.52	0.28	0.25	0.19	0.51	mile
Percentage of park area	6.30	11.26	9.04	8.06	1.26	%
Distance to Parks	0.92	0.35	0.51	0.80	2.46	mile
Distance to Sewers	0.33	0.68	1.64	4.04	8.55	mile
Density of Houses	3.00	2.30	1.78	1.26	1.06	per acre
Age of Houses	73.70	55.42	35.96	22.93	26.91	years
Developed Low Density	14.15	21.50	31.42	7.92	1.97	%
Developed Medium Density	24.90	29.81	12.19	2.65	0.48	%
Developed High Density	27.33	18.72	4.64	1.38	0.11	%
Deciduous Forest	1.73	9.46	13.14	16.70	10.33	%
Evergreen Forest	0.20	1.22	1.54	2.55	0.77	%
Mixed Forest	0.02	0.12	0.17	0.22	0.08	%
Shrub	0.01	0.34	1.23	2.03	1.90	%
Pasture/Hay	0.15	1.24	10.24	18.97	19.40	%
Cultivated Crops	0.01	0.54	10.86	13.67	49.32	%

The spatial distribution of urban classes based on the environment, built-environment, and proximity factors showed a concentric pattern. The City-High Density class in the center occupied the core regions of Twin cities of Minneapolis and Saint Paul. The City-Medium Density encircled the first class and included cities namely Golden Valley, St. Louis Park, Bloomington, and Richfield. The Suburb class was a

cusp between more urban classes (City-High Density and City-Medium Density) and relatively rural classes (Outer Suburb 1 and Outer Suburb 2). Some of the important small cities and towns classified in the Suburb class were Coon Rapids, Shoreview, and White Bear Lake in the North, and Shakopee, Inner Grove Heights, Rosemount, and Lakeville in the south. From Outer Suburb1 class one can start to see relatively rural characteristics which extended further to the Outer Suburb 2. Figure 35 shows the spatial distribution of the urban classes.

Figure 35 Five urban landscape classes in Twin Cities Metropolitan Area



In the TCMA, 1111 infected dead birds, approximately 6000 mosquitoes from the infected mosquito pools, and 85 human cases were reported from the year 2002 to 2007 (MDH; MMCD). Overlay of infected dead birds, positive mosquito pools, and human cases (assumed to be reported at the centroid of zip codes) on the five urban landscape classes in TCMA depicted a strong spatial association with the City-High density and the City-Medium density classes consistently from 2002 to 2007. Further, the occurrences of WNV incidences decreased with increasing distance from the core urban areas TCMA (Figure 36).

Figure 36 Spatial "Overlay" of West Nile virus incidences on the derived urban landscape classes in Twin Cities Metropolitan Area, 2002 – 2007

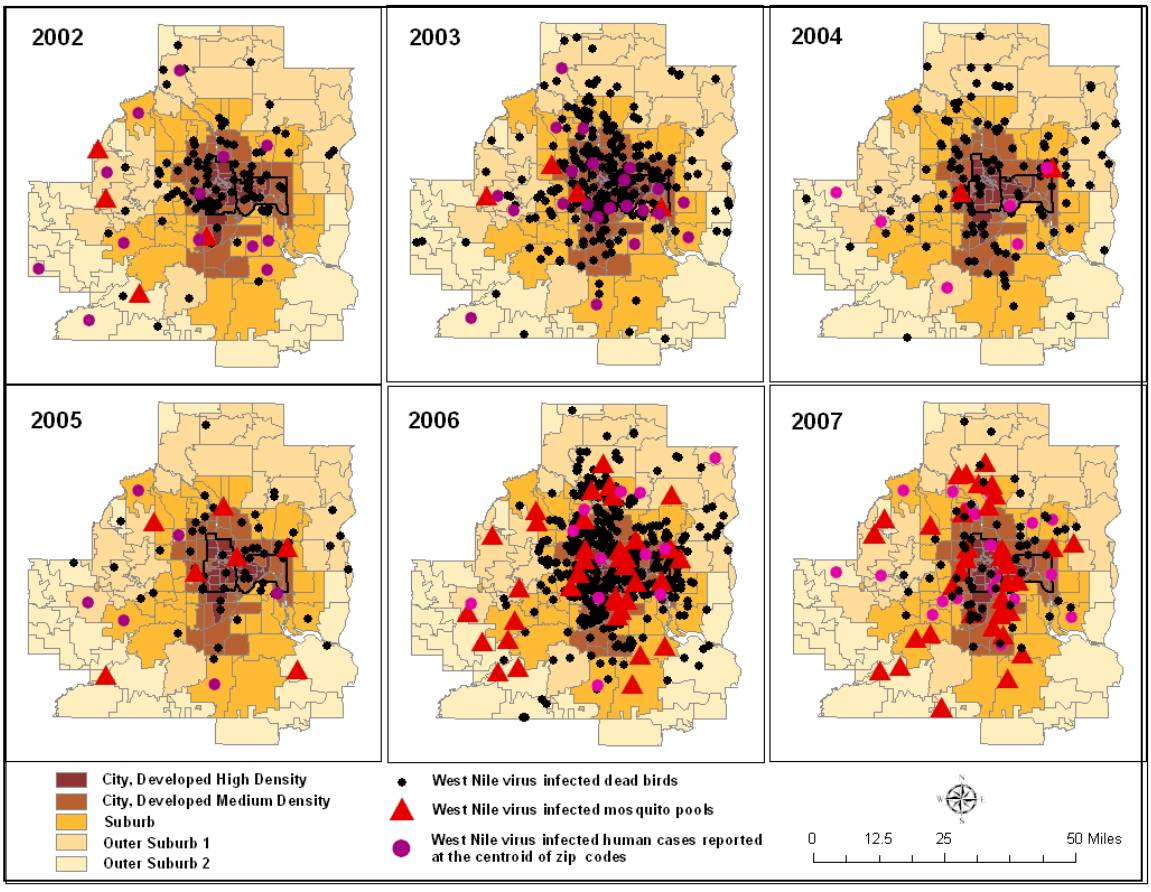


Table 24 Twin Cities Metropolitan urban classes with West Nile virus case rates among reported infected dead birds, positive mosquito pools, and humans from 2002-2006

Urban Class	Description	Number of Zip codes	Population	WNV Dead Birds per 100k	Mosquitoes in WNV infected pool per 100k	WNV Human cases per 100K
1	City-High Density	27	545267	912.89	88204.06	9.19
2	City-Medium Density	36	842625	1956.95	66162.71	11.12
3	Suburban	46	928789	1071.54	29440.40	6.18
4	Outer Suburb 1	28	360382	809.12	10127.02	2.85
5	Outer Suburb 2	22	152398	706.39	5085.07	2.91

The highest incidence rate of WNV infected dead birds was found in City-Medium Density class (1956.95) per 100,000 people. This was followed by the Suburban and the City-High density classes. The rate of infected dead bird reporting declined significantly in the outer suburb classes (Table 24). In the case of vector population, the City-High density urban class had the highest incidence rate of mosquitoes tested from the infected mosquito pools (88202.06 per 100,000 people). This rate declined further as we moved outwards from the core cities of Minneapolis and Saint Paul to the outer suburbs. The Outer Suburb 2 had the lowest vector incidence rate of 5085.07. Following the similar trend of WNV illness among birds and mosquitoes, the highest incidence rate among humans was reported in the City-Medium Density class, with 11.12 cases per 100,000 people. This was four times higher than the lowest rate of 2.85 cases found in the Outer Suburb 2 class and it was almost two times the rate found in the Suburb class (6.18). ANOVA tests for differences in zip code mean values further demonstrated the variation in incidence rate of mosquitoes among the urban classes. The mean values of City-High and City Medium density classes were significantly higher than the Outer Suburb 2 class (reference class) with p-values as

0.0269 and 0.0042 respectively (Table 25)

Table 25 ANOVA results of incidence rate of mosquitoes

Urban Classes	Estimate	T-value	P-value	Sig
City-High Density	1838.61	2.235	0.0269	*
City-Medium Density	3035.70	2.907	0.0042	**
Suburb	743.85	0.203	0.8393	
Outer Suburb 1	130.50	0.126	0.8999	
Outer Suburb 2 +				

F-statistic: 3.535 on 4 and 154 DF, p-value: 0.00863
+ Reference Class
* 95% significance level
** 99% significance level

5.4 Discussion

The three-step methodology of principal component and cluster analysis with 40 variables divided the TCMA into five concentric rings of urban classes. The City-High Density class was in the center, including the cities of Minneapolis and Saint Paul. This was encircled by other class of City-Medium Density, Suburb, Outer Suburb 1, and finally Outer Suburb 2. It is typical to assume that areas which are more natural would provide potential habitats for mosquitoes and are usually found away from urban areas. However, the degree to which an area is natural is not understood along a linear transect outward from the urban core. This generalization is often simplistic. For example in the City-High and City-Medium density classes, along with significant presence of built area, there were also natural areas in the form of lakes, parks, different wetlands, golf courses, trails, older residential and commercial buildings, and wedges along old

transportation routes. In addition, some of the features of built-environment namely urban storm water catch basins and ponds, construction sites, stock pile of abandoned tires, swimming pools in the backyards of residential houses can also provide potential breeding grounds for mosquitoes and increase the risk of WNV transmission. Such composition of urban landscape with natural and man-made features, typically affected by past land use and planning, creates an urban heterogeneous environment suitable for mosquito habitats. This is especially important in case of the TCMA, because two out of four WNV carrying vectors, *Culex restuans* and *Culex pipiens* are predominantly urban mosquitoes.

In a typical WNV transmission cycle, infected mosquitoes are a necessary prerequisite for human infections and therefore it is important to investigate whether the spatial variability of mosquito infection shows the same patterns as those found in human cases in the TCMA. The ESDA showed a strong spatial clustering of WNV infected dead birds, positive mosquito pools, and human cases in the urban/suburban areas (Figure 36). The incidence rate of human cases per 100,000 people was highest in the City-Medium Density class followed by City-High Density and Suburb class. City-High density class showed highest rate of mosquitoes tested in the infected mosquito pools. The difference in incidence rates for the mosquitoes and humans between the City-High and City-Medium density was not large. For example, the difference between human incidence rates was lower than two cases per 100,000 people. In addition, the ANOVA results statistically demonstrated that the means of mosquito incidence rates for zip codes in City-Medium and City-High density urban classes were significantly different than the other classes. Thus, these results ranging from exploratory to confirmatory analysis provided evidences that both the incidence rates of infected mosquitoes and human cases followed similar spatial pattern. The incidence rates were higher in the urban heterogeneous environment of the City-High and the City-Medium Density classes and much lower in the more natural environment of outer suburbs. This also confirmed the hypothesis (Chapter 4) that the nature of WNV transmission in TCMA is urban/suburban centric.

The urban landscape features dominant in the City-High density and the City-Medium density classes are as follows. Density of urban catch basins and storm water ponds, built primarily to accumulate polluted urban run-off, emerged both as a predominant feature of urban structure and was strongly associated with WNV occurrences. The rate of water drainage and the presence of organic matter in catch basins affect the breeding conditions for *Culex* mosquitoes. This is especially true in times of drought when the water stands for a longer period of time (Shaman et al. 2005). Housing characteristics namely density of buildings (both residential and commercial) and age of houses played important role in the viral activity of the virus. This finding coordinated with the findings of the study based in Chicago and Detroit (Ruiz et al. 2007). Specifically, the positive association between houses built during the post World War II period from 1940 – 1960 and WNV illness in the City-Medium Density class of TCMA and Inner Suburb class of Chicago deserved special mention. It was possible that during that time due to sudden increase in population and demand for housing, less attention was given to the surrounding drainage structure, physiography, and soil characteristics. This often resulted in basement flooding which was quickly mitigated by building catch basins very near to the houses. These catch basins, with time accumulated water and decaying plant residues and if left untreated, could likely provide suitable breeding grounds for *Culex* mosquitoes. This will further increase the risk of WNV (Ruiz et al. 2007).

In the City-Medium Density class, patches of natural areas were also observed in the form of parks, lakes, shallow marsh, wooded swamps, bogs, vegetation in the form of trees, shrubs, grassy alleys, trails, (natural and unpaved), and golf courses. The dynamics of virus circulation between birds and mosquitoes and their interaction with such natural habitats in urban places are fundamental in understanding the transmission of the virus. Among the land cover variables from NLCD data set, developed-high density, developed-medium density, developed-low density, pasture/hay, and cultivated crops contributed to the City-High density, City-Medium Density, Suburb, Outer Suburb 1, and Outer Suburb 2 respectively.

Age of houses, especially “older houses” of 50 to 60 years old was a strong variable in this analysis. Given the possibility that in low-income neighborhoods, older houses are not well maintained, future analysis might include the interaction effect of income and age of houses. Additional variables differentiating residential, commercial, and industrial use, impervious surface, and knowledge of soil characteristics can also be helpful. These variables can provide a better understanding of natural open space and impervious surface, which would be helpful in identification of potential mosquito habitats.

This study had the following similarities and dissimilarities with the study based in Chicago and Detroit (Ruiz et al. 2007). Among the similarities, the landscape characterization of urban classes in TCMA, Chicago, and Detroit was conspicuous. In all the three metropolitan areas, the derived urban classes showed a typical concentric pattern of urban growth with urban core or Central Business District in the center. Based on the NLCD land cover categories and host of other environment and built-environment features, the TCMA was divided into five urban classes of City-High Density, City-Medium Density, Suburb, Outer Suburb 1, and Outer Suburb 2. Similarly, in Chicago the five classes were City-Low income, City-High Income, Inner Suburb, Outer Suburb, and Urban N-man’s. The classes in Detroit were City-Low Income, Inner Suburb I, Inner Suburb II, Outer Suburb I, and Outer Suburb II. The highest incidence rate of WNV infected human cases per 100,000 people was found in the Inner Suburb class in Chicago (22.78), Inner Suburb II class in Detroit (22.76), and in the City-Medium Density class in TCMA (11.12). In the TCMA, higher incidence rates were found in the urban areas and the rates declined significantly in the relatively rural landscape of outer suburb classes. Thus, all the three metropolitan areas showed evidence of urban centric nature of WNV transmission and amplification.

Among the various urban features, age of housing, especially belonging to the 1940-60’s period, was the dominant factor in the urban classes with highest incidence rate in the TCMA, Chicago, and Detroit. In addition these classes also showed patches of natural areas in highly built-up area. This indicated that the composition of natural

and built-environment features in an urban landscape is conducive for WNV dynamics and transmission of the virus from one component to another.

As for the dissimilarities, this study included more variables in the categories of built-environment and proximity variables, which were not considered in the other studies. For example, density of urban catch basins and storm water ponds, distance to 8 different wetland types, trails, parks, lakes, and golf courses were new variables incorporated in this study to capture the detailed landscape features of heterogeneous urban environment conducive for WNV transmission. A major contribution of this particular study was that along with human incidence rate, this work also investigated the association of the derived urban classes with WNV infection in mosquitoes. Ruiz *et al.* indicated in their study that the occurrence of WNV infection among mosquitoes is a necessary precursor for human infections and therefore it is important to explore whether the spatial variability of association of mosquito infection follows the same pattern as found in human illness (Ruiz et al. 2007). In TCMA, exploratory spatial analysis showed a strong spatial clustering of WNV infected mosquito pools in the urban areas (Figure 36). The incidence rate of mosquitoes in the pools tested positive was highest in the City-High Density class followed by the City-Medium Density and the Suburb classes (Table 24). In addition, the ANOVA analysis further proved that the mean of incidence rate of mosquitoes was significantly different in the City-High and the Medium classes than the other classes (Table 25). Similar results were found for WNV human incidence rate which was highest in the City-Medium Density class, followed by City-High Density and Suburb classes. Therefore, this study empirically showed similar pattern of spatial variability of WNV infections among mosquitoes and humans in the TCMA, which was only hypothesized in other studies.

5.5 Conclusion

Following a three-step methodology, the urban landscape classes derived from the combination of environment and built-environment variables, indicated positive association with WNV illness in the TCMA. The specific urban features that contributed to the viral activities of WNV were catch basins, housing density, age of houses, and proximity to lakes, swamps, bogs, and parks present within the urban areas. The City-High density class, including the core urban areas of Minneapolis and Saint Paul, and City-Medium density class, covering the immediate suburbs, reported highest incidence rates of infected mosquitoes and human cases. These results indicated that the Twin Cities urban landscape, a combination of natural and man-made features, created heterogeneous environments suitable for mosquito habitats. This is especially important because two out of four WNV carrying vectors, *Culex restuans* and *Culex pipiens* are predominantly urban mosquitoes. These research findings were in line with the urban centric nature of WNV transmission in Chicago and Detroit and therefore strengthened the general hypothesis that WNV infection has exhibited strong spatial clustering in the urban areas of Midwestern United States. The associations with the risk factors will further help in developing hypothesis to understand the causal factors of WNV occurrences. In addition, the derived urban landscape classes can provide a basis for the selection of field sites for mosquito collection, testing, and treatment. This could be very useful for personnel at the MMCD for selecting sites for setting up mosquito traps and collection. Finally this study also contributes to the broader research question in the field of health geography and public health, of how the heterogeneous urban landscape affects human health and disease patterns.

In the following chapters, the association between the selected environment, built-environment, and proximity risk factors and the occurrences of WNV incidences are examined rigorously in a modeling framework.

6. Chapter 6: Modeling the dynamics of West Nile virus in the Twin Cities Metropolitan Area of Minnesota

6.1 Background

The modeling techniques to predict and understand the relationships between WNV occurrences and the hypothesized natural and anthropogenic risk factors can be broadly divided into three groups - linear, nonlinear, and spatial.

Ordinary Least Square (OLS) is the most common statistical technique used in the linear category (Yiannakoulias et al. 2006; Ezenwa et al. 2007; Mongoh et al. 2007; Pradier et al. 2008). Even though linear models are simple to interpret, they suffer from several disadvantages especially when dealing with complex health outcome. In reality, the transmission of a multi-host infection, such as WNV, is complex and therefore it is challenging for linear models to account for nonlinearity of the relationships between disease outcome and risk factors. Moreover, the inflexibility of OLS models due to rigid assumptions (normality, independent observations, and homoskedasticity) makes it inappropriate to model complex phenomena.

Nonlinear models employed in the WNV research can be divided into two broad groups, intrinsically linear and intrinsically nonlinear models. The former class of models can be transformed to a *linear* function and subsequently analyzed using linear models and thus are not entirely nonlinear. Examples of these types of methods include Poisson, logit, and probit regression models. However, in the intrinsically nonlinear

models, nonlinear form of a model cannot be transformed to a linear form. Examples of this type of models include the general growth model, maximum likelihood, and neural network algorithms. There are several examples of previous attempts to model WNV disease occurrence by logistic (Ruiz et al. 2004; Shaman et al. 2005; Diuk-Wasser et al. 2006; Gibbs et al. 2006; Murray et al. 2006; Leblond et al. 2007; Brown et al. 2008; LaBeaud et al. 2008) and Poisson regressions (Roberts and Foppa 2006; Yiannakoulias et al. 2006). There is no doubt that these studies contributed significantly to the limited knowledge of a rapidly spreading virus but can be improved further by using sophisticated *nonlinear* models. For both Logistic and Poisson regression models, a user has to specify an *a priori* functional form or in other words a statistical distribution such as binomial or Poisson distribution between the response variable (WNV infection) and the predictors (risk factors). However, given the novelty and complexities of the transmission characteristics of WNV in the western world, the interaction of avian-mosquito-human habitats coupled with environmental, built-environment, and sociodemographic risk factors, the assumption of a predefined distribution is simplistic. Therefore it is difficult for these intrinsically linear models to incorporate the nonlinearity between the disease occurrence and the risk factors.

The most common spatial models used in the WNV research are Generalized Linear Mixed Models (GLMM) (Yiannakoulias et al. 2006; Yiannakoulias and Svenson 2007; Johnson 2008). The main advantage of the spatial models is their ability to account for spatial autocorrelation, especially when multi-scale environmental, climatic, and social risk factors are included in the models as predictors. However, this group of model also suffers from similar weaknesses as that of linear and logistic regression models that are first, assumption of a predefined relationship between response and predictor variables and second, lack of nonlinearity.

The goal of this chapter is to develop computational neural network (CNN) models to predict the incidences of WNV infected dead birds by capturing the complex or *nonlinear* relationships of the hypothesized risk factors in the TCMA. There are several advantages of using neural network algorithms in understanding a complex

health outcome, such as WNV, over the linear and intrinsically linear models (Logistic and Poisson). The advantages are as follows: 1) ability to capture complex relationships between risk factors and disease occurrence 2) good predictive capability, 3) does not require to specify *a priori* distribution because it learns the relationships between the input and the output from the dataset itself, 4) no rigid assumptions of normality and homoskedasticity. Though CNNs have been successfully employed as predictive tools in the information and natural sciences, they have received less attention in social and health sciences to predict the spread of infectious diseases. The results obtained from the CNN model in this work are also compared to that of the OLS model with the same specifications to assess the higher predictive capabilities of CNN models.

This chapter is organized as follows. The section 6.2 discusses the underlying details of the various modeling and optimization techniques used in this work, section 6.3 describes the data and procedure involved in developing a nonlinear WNV analysis model, section 6.4 reports the results, and finally section 6.5 concludes the chapter.

6.2 Description of Modeling and Optimization Techniques

This section discusses the underlying details of the various modeling and optimization techniques used in this work.

6.2.1 Multiple Linear Regression

A linear relationship between an observed value (i.e. response value) and its independent variables can be modeled by

Equation 6.1 $$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i \qquad i = 1, 2, \ldots\ldots, n$$

where y_i is the response and $x_{i1}, x_{i2}, x_{i3}, \ldots\ldots, x_{ip}$ are the independent variables for the i^{th} observation and n is the number of observations. $\beta_0, \beta_1, \ldots\ldots, \beta_p$ are parameters that are to be estimated. ε_i is the error term and is assumed to be a normally distributed random variable. It is also assumed that the standard deviation of the error term is constant and it does not depend on the "x" value or the independent variables. This assumption is called homoscedasticity.

Multiple regression is a technique by which y_i and $\beta_0, \beta_1, \ldots\ldots, \beta_p$ can be estimated. Thus Equation 6.1 may be written as

Equation 6.2 $$\widehat{y_i} = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots\ldots + b_p x_{ip}$$

where $b_0 + b_1 + b_2 + \ldots\ldots + b_p$ are the estimated values of the parameter in Equation 6.1. The most popular algorithm to estimate the parameters is the OLS method. The best fitted OLS model is one which minimizes the square of the error term between the predicted response, $\widehat{y_i}$ and the observed response, y_i .

After estimating the model parameters, the goodness of fit or the model quality can be ascertained in a number of ways. Two common measures of model quality are the R^2 value and the root mean square error (RMSE). The R^2 is also known as the Pearson Coefficient and its value ranges from -1 to +1. The RMSE is defined as

Equation 6.3 $$RMSE = \sqrt{\frac{1}{n}\sum(y_i - \widehat{y_i})^2}$$

Typically good models are characterized by high value of R^2 and low value of RMSE. Other indicator of model quality include *F* – statistic.

Individual predictions can be examined by a variety of methods. The most effective and simplest method is to plot residuals (the difference between the observed

value and the predicted value) verses the index number of the observed value. A good model will show a normal random distribution of residuals. Distinct patterns (such as upward or downward trends) are indicative of heteroscedasiticity, i.e. presence of unequal variance, and the dataset must be reexamined. Patterns showing concentrated pockets of high or low values of residuals signify the presence of spatial autocorrelation, i.e., the observations are not independent. In such a situation, efforts should be made to correct for spatial autocorrelation or use other regression techniques which do not have an assumption of independent observations. The residuals can also be examined by making a normal probability plot (Q-Q plot). If the residuals do have a normal random distribution, this plot will be a straight line. Deviations from this will indicate the presence of outliers. In case of residuals obtained from Student-t test, residuals typically lying above 2.0 or below -2.0 are identified as outliers. Outliers can also be determined by the use of other diagnostics such as Cook's distance (Cook and Weisberg 1999) and Mahalanobis distance (Schinka et al. 2003). Once outliers are detected the model can be regenerated by excluding those observations from the data set.

6.2.2 Neural Networks

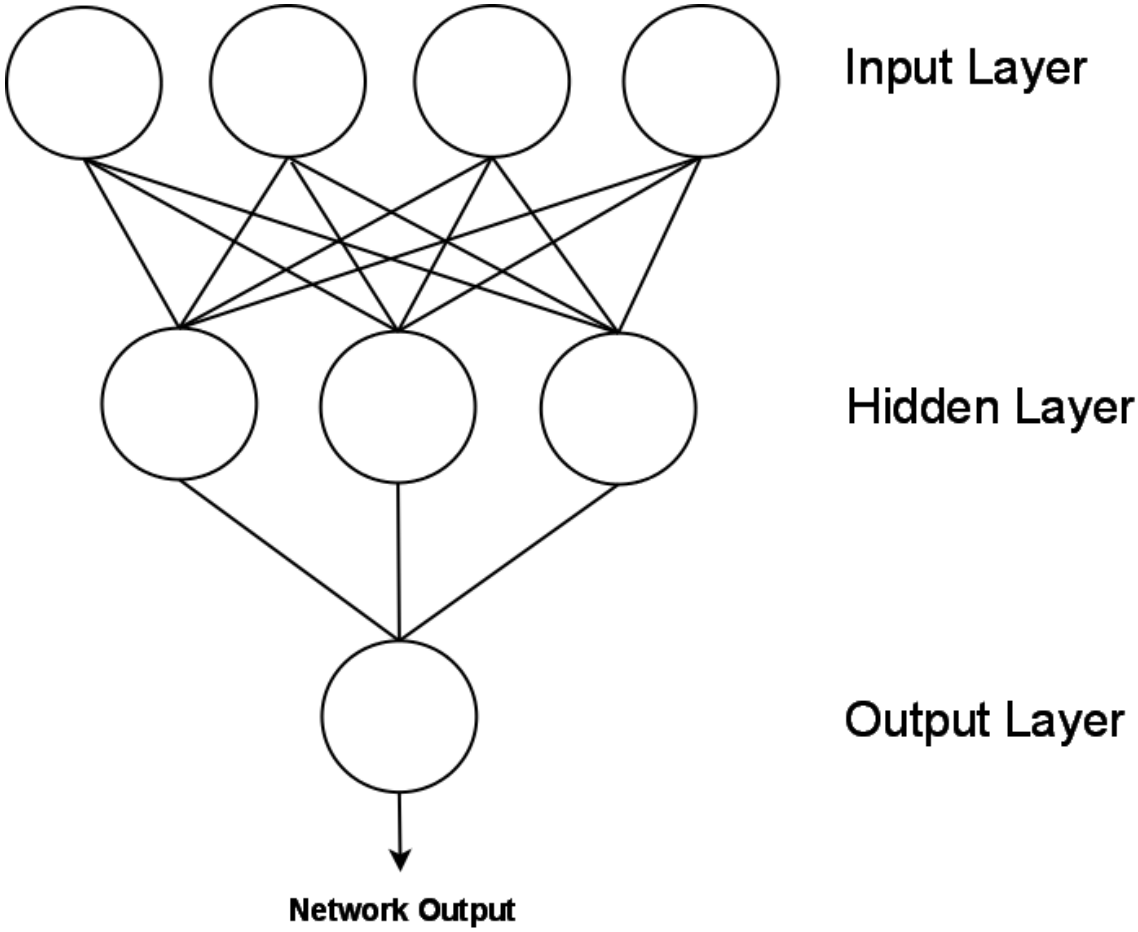
In this work I mainly focus on the use of Computational Neural Network (CNN) algorithms to model the dynamics of WNV in the TCMA. According to the classification scheme described in the section 6.1, CNN belongs to the intrinsically nonlinear modeling category, in which the functional form cannot be transformed to a linear form. CNN essentially attempts to mimic the behavior of a human brain and thus a fundamental characteristic of these algorithms is the ability to *learn* the relationships present within a dataset. It resembles the behavior of a human brain in the following aspects: 1) the basic unit of operation in a neural network is *neuron*; 2) knowledge is acquired by the network from the environment (dataset) through a learning process, and 3) interneuron connection strengths, known as synaptic weights, are used to store the acquired knowledge (Guha 2005c). As mentioned above, CNNs are

a specific class of nonlinear models and they differ from traditional nonlinear models in the following ways: 1) CNNs do not require specifying an *a priori* distributional form such logistic or Poisson distributions, and 2) CNNs do not represent the relationships within a dataset in an explicit functional form, instead the relationships in the datasets are encoded by a set of connections between units termed *neurons*.

Haykin (2001) described a neural network as "an extensive parallel distributed processor made up of simple processing units, which has a natural propensity and ability to store knowledge through learning and making it available for further use". There are a large variety of neural network algorithms namely feed forward, back propagation learning, probabilistic, radial basic function, self-organizing map, and structural learning and forgetting algorithm (Haykin 2001). Among these algorithms, I have used feed-forward neural network algorithm for this particular work mainly for two reasons. First, feed forward algorithms are simpler than the other algorithms mentioned above. Second, the techniques used to interpret the CNN models in this work are *only* developed for feed forward neural network algorithms (Guha 2005c).

The structure of a feed-forward neural network model includes three layers, which are fully connected. The first layer is termed the input layer and each neuron in this layer corresponds to a predictor variable in the model. The second layer is termed the hidden layer and is responsible for nonlinearly combining the values of the predictor variables in the model. The final layer is termed as the output layer whose output is the predicted value of the response variable. The term fully-connected indicates that all the neurons in a given layer are connected to all the neurons in the next layer. The Figure 37 shows the structure of a fully connected 3-layer neural network.

Figure 37 A schematic diagram of a 3-layer, fully connected feed-forward neural network



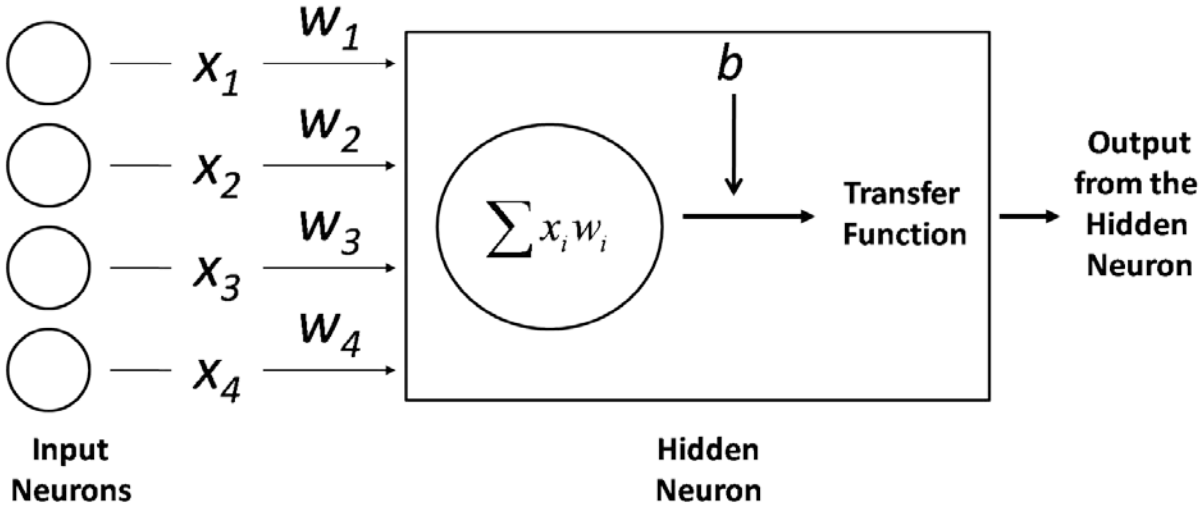
The mechanism of a neural network is based on the neurons. At each layer, a neuron accepts input values and weights, associated with the nonlinear functions in the preceding layer and these values are then transferred to the next layer by a *transfer function*. The main advantage of a neural network is that the transfer function is

generally nonlinear in nature. A number of transfer functions are described in the literature (Haykin 2001) and the implementation used in this study applies a sigmoidal function given by the following formula

Equation 6.4 $$o = \frac{1}{1 + \exp\left(-\sum x_i w_i + b\right)}$$

where O is the output of the neuron, x_i is the output value of the i^{th} neuron in the preceding layer, w_i is the weight for the connection between this neuron and the i^{th} neuron in the preceding layer and b is the value of the bias term (Guha 2005a). Figure 38 is a graphical representation of a single hidden neuron receiving input values and weights from the input neurons and passing these values to the next layer through a transfer function.

Figure 38 A more detailed view of a single layer hidden neuron



Note: Adapted from the figure in source: *Guha, R. 2005a. Methods to improve the Reliability, Validity, and Interpretability of QSAR Models in Chemistry, Pennsylvania State University, State College, PA.*

In the Figure 38, x_i's represent the output value of the neurons in the preceding layer, w_i's correspond to the weights for the connections between this neuron and those in the preceding layer, and *b* represents the bias term for this neuron. In case of a 3-layer network, weights between the input and the hidden layer neurons allow the network to be configured so that the important predictor variables will have greater contributions to the hidden layer neurons. The weights and biases, transferring from one layer to another, allow the network to *learn* the features present in the dataset and therefore allowing the CNN to make accurate predictions.

The next important step in the neural network algorithm is to obtain an optimal set of weights and biases, which will allow the network to make better predictions. Typically the number of weights and biases is specified by the structure of the neural network, i,e, by the number of input (predictor variables) and hidden layer neurons. Increasing the number of hidden neurons will improve prediction but with a cost of over fitting (Guha 2005c). One rule of thumb to determine the appropriate number of weights and biases is that the total number of parameters should be less than half the size of the training set used to build the model, that is, (Livingstone and Manallack 1993)

Equation 6.5 $$n_I n_H + 2n_H + n_O \leq \frac{n_{TSET}}{2}$$

where n_I, n_H, and n_O are the number of input, hidden, and output neurons respectively and n_{TSET} is the number of observations in the training set.

Next, a CNN model is initialized to train the network. To do so, I used the *nnet* function from the package nnet in the R statistical software programming language (http://cran.r-project.org/). In the *nnet* function, the network is initialized with a set of weights and biases using a random number generator and the network is trained by BFGS optimization algorithm. An important feature of the training phase is that it is supervised. That is, to train the network, observed values of the training set are required. It is also important to note that since the network is initialized using random numbers,

multiple runs of the nnet function can result in slightly different sets of optimal weights and biases. As a result, good practice suggests that the final result is obtained using an ensemble of models obtained from multiple runs of the nnet function (Guha 2005a). Once the model has been trained, its quality can be evaluated by noting the R^2 and the RMSE values.

6.2.3 Cross-Validation

Once the model is built, cross-validating is the next important step. Model validation can be conducted using several cross-validation techniques both internally and externally. The cross-validation is termed internal when a technique is used as a criterion in building a model or choosing a best model from an ensemble of models. On the other hand, for an external cross-validation technique, new or external prediction datasets are used. For this work, Leave-one-out cross validation or LOO method is used as an internal technique (Golbraikh and Tropsha 2002). The Q^2 , obtained by using the LOO cross-validation procedure is an alternative to R^2. That is, the neural network model with the same structure (same number of input variables, hidden neurons, and output neuron) is generated using the whole dataset excluding one point. The response value for this point is then predicted using the model and this procedure is repeated for all the points in the dataset. The R^2 for these predictions is denoted by Q^2. Typically if the R^2 and Q^2 are in the range of 10-15 percent, the model is considered good with high predictive ability and generalizability. For external cross-validation methods, the RMSE and the R^2 values obtained from the new datasets are noted. Ideally one would expect higher R^2 and RMSE values usually within 10 percent range of the original response variable.

6.2.4 Optimization Method ~ Genetic Algorithm

In social and health sciences, statistical and mathematical modeling begin with considering a large number of predictor variables, especially when investigating a

multi-host health outcome (Lyme disease, WNV, malaria, SARS, etc), land use /land cover analysis, or other human-environment interactions. The pool of predictor or explanatory variables often includes multidimensional factors such as environmental, built-environment, social, economic, and demographic variables. Even though including a large number of predictor variables to some extent prevents issues associated to missing variables, it often leads to *over fitting*. This is especially true when the sample size or *n* is small. Following the theory of parsimony and minimizing overfitting, the goal here is to include the minimum number of predictors that will explain the maximum variation of the response variable. Thus, we must objectively select a *good* or *best* subset of predictor variables. Here the degree of goodness of fit will be evaluated by the modeling technique used as well as the results of cross-validation techniques.

Several statistical methods exist which perform variable selection such as stepwise regression, backward, and forward selection. However these methods are generally restricted to linear regression methods or nonlinear regressions like logistic regression, which can be linearized. However, a more important reason for not considering these methods is due to number of drawbacks including falsely calculated narrow confidence intervals (Altman and Anderson 1989), incorrect p-values, biased regression coefficients, and problems with multicollinearity and so on (Guha 2005a). Furthermore the forward and backward selection algorithms by their nature will ignore certain combinations of variables, since in the former case variables are based on the current subset that has already been selected and in the latter case variables removed from consideration are not considered again. This will significantly affect in modeling a phenomenon where interrelationships and combinations of predictor variables is important. Therefore I used an alternative optimization algorithm to carry out variable selection. Among the several deterministic and stochastic optimization methods available, I have used Genetic algorithms (GA) to identify a best subset of predictor variables. The following paragraphs describe the principles underlying a GA.

A genetic algorithm is a stochastic class of optimization techniques known as evolutionary algorithms, which utilizes the concepts of biological evolution such as

reproduction, crossover, and mutation to develop efficient optimization strategies (Forrest 1993). As a result much of the terminologies used in GA are adapted from the field of biological evolution. The typical terminologies used in a GA are population, individual, chromosome, fitness value, crossover, mutation, and child population. These terminologies are defined sequentially in the following paragraphs. When using GA for variable selection, a *chromosome* is defined as a subset of predictor variables (of user specified size) chosen from the entire predictor pool that is being searched. An *individual* is comprised of a chromosome and an associated *fitness value.* A *population* is defined as a collection of individuals.

The first step in a GA is to initialize the population by randomly generating a user specified number (usually 50 to 60, though this is dependent on the size of the predictor subset desired) and size of predictor subsets (Guha 2005a). Each subset is then used to build a model (which can be either linear regression model or a CNN model) and the RMSE for each model determines the fitness of the individual. The population is then ranked based on the fitness value for each individual.

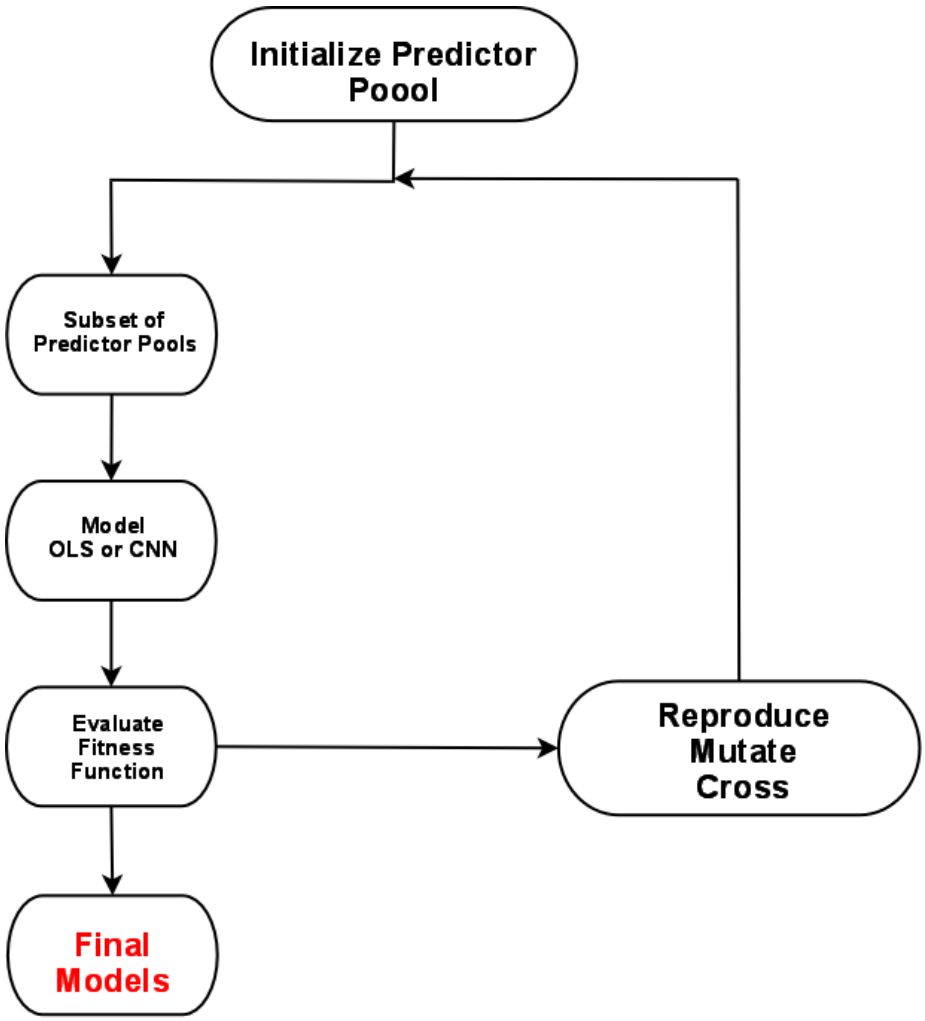
The next step is to create a child population and the steps are listed below.

1. A mating list is created, which is of the same size as the current population.
2. Those individuals with fitness value (RMSE) greater than population average are automatically placed in the mating list.
3. The remaining positions are filled by using a roulette wheel selection procedure to select individuals from the current population (Guha 2005a)
4. Finally a child population is generated by randomly selecting two individuals from the mating list and performing genetic operations such as *crossover* and *mutation*.

Crossover involves the swapping of portions of the chromosomes of a pair of individual (Forrest 1993). The goal of the crossover function is to generate a new individual with good features of their parent population. In other words, if both the parents have a high fitness values this implies that parts of their chromosomes (i.e. predictor variables) are responsible for their fitness. Thus by combining a portion of the

chromosomes of fit parents, children are produced with equal or better fitness. The second genetic function is mutation and is performed on individuals in the child population. Mirroring the low frequency of mutation in biological evolution, mutation is not performed on all the individuals in a population but carried out only 5 percent of the time. In a GA, the mutation is performed by randomly changing a part of the chromosome (i.e., randomly replacing one of the predictor variable) of an individual and its main function is to maintain diversity of the population. Figure 39 shows a flow chart of a generic GA.

Figure 39 A flow chart describing the steps involved in a Genetic Algorithm



With the application of these operations, a second child population is generated and their individuals are again ranked based on their fitness value. The second generation population is then generated by randomly selecting individuals from the top 50 percent of the child population and the whole process is repeated for a user specified number of iterations (typically 1000). Finally top ranked individuals (i.e. top ranked predictor subsets) with their fitness value (RMSE) are reported to the user.

6.3 Data and the proposed WNV Analysis Model

6.3.1 Data

The WNV incidence data for the year 2006 was used to build the model. Infected dead birds aggregated by zip codes was the response variable and the predictor variables belonged to either of the four risk factor categories of environmental, built-environment, proximity, and vector control programs. The data from the year 2006 was chosen mainly because the number of infected dead bird was highest (479) in this year. After checking for collinearity, missing values, and data ranges (variation), 32 potential risk factors of WNV were included in the model. They were maximum daily temperature, daily precipitation, and land cover variables in the environmental category. Built environment factors mainly included density of catch basins, density of ditches, area of parks (open green space), housing density, and housing age. Proximity to features such as lakes, wastewater discharge points, golf courses, trails, shrub swamps, wooded swamp, and bogs were also considered. The variables in the final category of vector control policies were frequency and percentage of public land survey (PLS) units treated for larvicide and adulicide. All the variables were processed and aggregated at the census zip code level (refer Chapter 2 for details). Table 26 shows the summary statistics of all the variables included in the model.

Table 26 Descriptive summary of variables in the model

Variables	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Unit
Response Variable							
WNV Infected Dead Birds	0	0	2	3	5	17	count
Predictor Variables							
Daily Max Temperature	75.39	80.31	82.93	79.52	84.48	98	F
Daily Precipitation	0	0	0.08804	0.1494	0.2505	1	inches
Developed, Open Space	0	4.057	7.647	8.92	13.47	24.72	%
Developed, Low Density	0.6151	4.405	17.03	18.8	30.06	64.28	%
Developed, Medium Density	0	2.071	12.52	12.77	20.09	39.11	%
Developed, High Density	0	0.5473	4.291	9.12	11.53	81.24	%
Shrub/Scrub	0	0	0.4003	1.054	1.704	6.838	%
Pasture/Hay	0	0.363	3.565	9.293	17.23	42.94	%
Cultivated Crops	0	0	1.31	12.5	19.33	76.74	%
Density of Catch Basins (Wet)	0	4.026	23.18	43.94	47.44	383.1	sq mile
Density of Catch Basins (Dry)	0	22.85	171.5	200.4	310.7	740.1	sq mile
Density of ditches	0	0	0	0.1238	0.1264	1.145	sq mile
Housing Density	1	1.164	1.73	1.914	2.351	5.029	acre
Age of houses	0	25	37	41.21	52	90	years
Density of bike paths	0	0	0.4244	0.8954	1.28	4.682	sq mile
Percentage Area of Parks	0	2.783	6.508	7.827	10.7	41.61	%
Flooded Basins and Flat (D)	177.3	364.8	503.6	602.9	687.9	2394	miles
Inland Fresh Meadow (D)	108.2	290	440.2	586	654.4	2143	miles
Inland Shallow Marsh (D)	167.2	379	530.3	792.2	985.2	2990	miles
Inland Deep Fresh Marsh (D)	136.2	295.9	439.8	555.4	683.4	1854	miles
Inland Fresh Open Water (D)	420.7	747.9	1067	1603	1941	13620	miles
Shrub Swamps (D)	558.1	2243	4242	5008	7092	17250	miles
Wooded Swamp (D)	1359	4873	8295	8977	12110	38800	miles
Bog (D)	661.4	2835	5655	6996	8849	25300	miles
Lake (D)	60.59	301.5	440.5	524.3	646.1	1872	miles
Park (D)	117.7	288.5	442.1	1147	1461	9027	miles
Water Discharge Points (D)	240.5	1017	1960	4334	5890	19930	miles
Golf Courses (D)	1202	2275	2897	3838	4206	21490	miles
Trail (Unpaved) (D)	138.5	1194	2639	5309	6912	35680	miles
Percentage of Larvicide Treat	1.639	59.51	92.86	76.9	100	100	%
Frequency of Larvicide App	2	15	28	31	45	83	count
Percent of Adulticide Treat	0	10.74	21.28	26.26	38.17	100	%
Frequency of Adulticide App	1	2	3	3	4	19	count

6.3.2 WNV analysis Model

For statistical analysis I have used R statistical programming language (http://cran.r-project.org/). The main functions and packages used to build the neural network model were *nnet* function from the library "nnet" and modified genetic algorithm *rbga* function from the package "genalg".

The steps involved in the model building are as follows:

1. **Choosing a subset of predictor variables**: The predictor pool consisting of 32 variables (Table 26) were passed to the modified *rbga* function. The number of variables (*N*), population size (100), and iterations (1000) were specified. The output was 100 possible subsets of predictor variables of size *N*, let us say $Vars_{1\ to\ 100}$.

2. **Ensemble of Neural Network Models**: Using the *nnet* function, 100 CNN models were built with $Vars_{1\ to\ 100}$ with user specified number of hidden neurons (*HN*). For each CNN model, the response variable was infected number of dead birds by zip codes and predictor variables were one of the subsets of $Vars_{1\ to\ 100}$. In addition, the R^2 (square of the correlation between observed *y* and predicted *y*) and the RMSE (Equation 6.3) values were calculated for each of the 100 models and stored for final model selection.

3. **Cross-Validation:** The Q^2 values for the 100 CNN models were calculated by the Leave One Out (LOO) cross-validation method (see section 6.2.3 for details). For each CNN model, the Q^2 value was compared with the R^2 value.

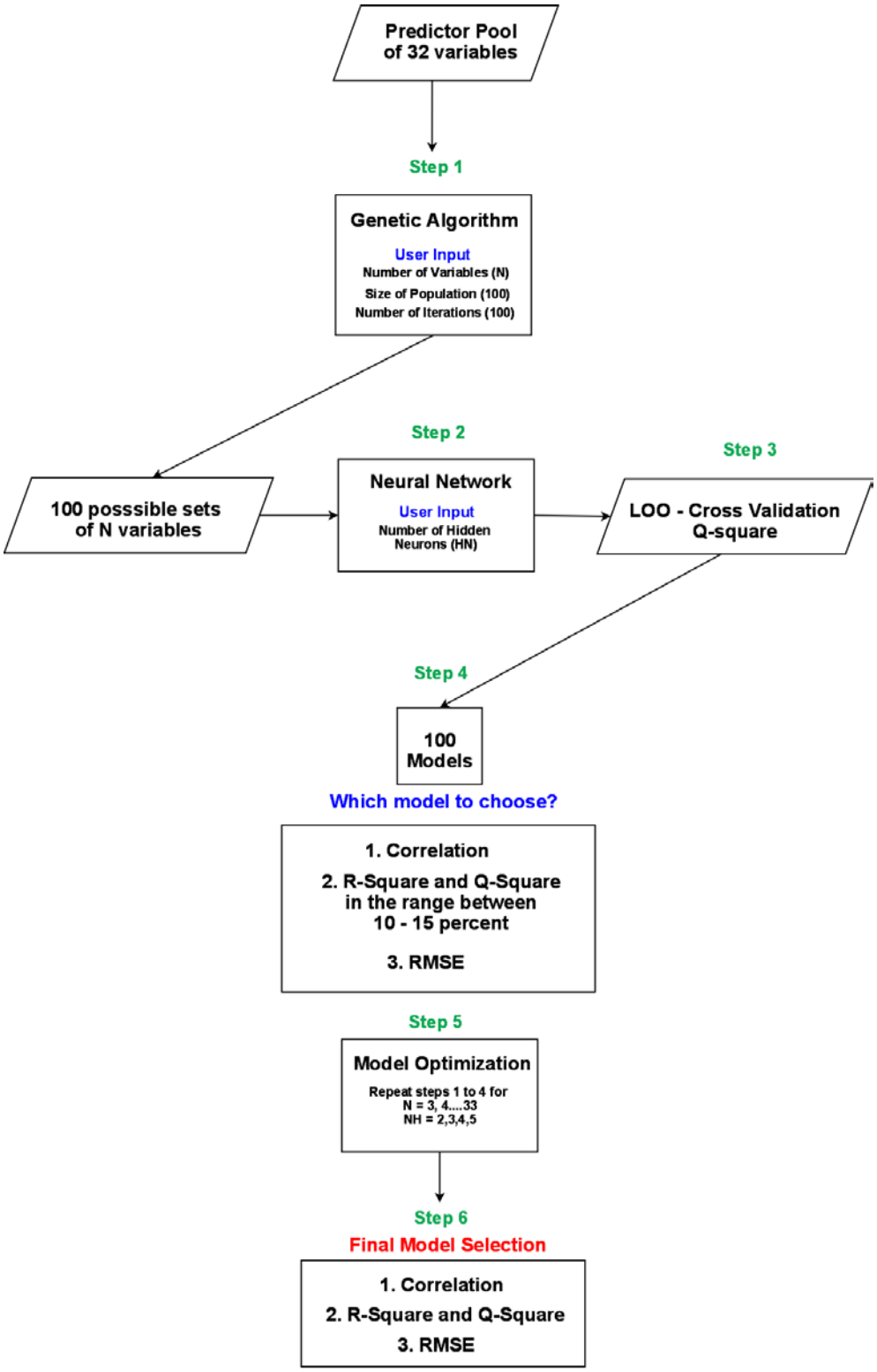
4. **Selection of best *N*-predictor model:** Among the 100 CNN models of *N* predictor variables, the best *N*-predictor model was chosen based on two criteria. These criteria were 1) low value of RMSE and 2) R^2 and Q^2 values were within the range of 10 – 15 percent.

5. **Model Optimization:** The steps from 1 to 4 are repeated for variable size $N = 3, \ldots..32$ and number of hidden neurons for $HN = 2, 3, \ldots...5$. Here I have restricted the number of hidden neurons to 5 because with 159 records (zip codes), increasing the hidden neurons beyond five will result in over fitting. After completion of all the

optimization runs, we obtained the best 3-predictor model, 4-predictor model, 5-predictor model and so on.

6. The Final Model Selection: The final model was chosen from these best *N*-predictor models by comparing their RMSE and the difference between the R^2 and Q^2 values. The final model had lowest difference between the R^2 and Q^2 values and relatively low RMSE value.

Figure 40. A flow chart describing the workflow of modeling and optimization techniques used to select the final model



6.4 Results

The steps to select the best nonlinear WNV analysis model described in the previous section are executed here. The architecture and statistics of the final model are discussed. In addition, to demonstrate the effectiveness of the proposed approach, the outputs from the CNN model are also compared with the results obtained from the OLS model with same model specifications.

6.4.1 Final Model Selection

The Figure 41 shows the change in RMSE values calculated from the CNN models with an additional increase of predictor variables (up to 14 variables) and an additional increase of hidden neurons (up to 5 hidden neurons). The number of variables is shown on the x axis and RMSE value on the y axis. The color coded lines denote models with different number of hidden neurons. In this figure I purposefully chose to report the RMSE values up to 14 variables because beyond N=14 the RMSE's were noisy with no distinct trend. Similar was the behavior of RMSE values with additional increase of hidden neuron beyond five. The first sharp decline in the RMSE value was identified at models with five input variables. After this point, with an additional increase of a risk factor, the RMSE values started to increase and showed inconsistencies with irregular *peaks* and *dips*. This overall trend of RMSE values was similar for all the models with hidden neurons from 2 to 5.

Figure 42 shows the difference between the R^2 and Q^2 values. Here number of variables is shown on the x axis, difference between R^2 and Q^2 on the y axis, and the colored lines denote models with increasing number of hidden neurons. The figure shows that the CNN model with two hidden layer neurons and five predictor variables had the lowest difference between them. Thus, based on the trend of RMSE values in Figure 41, and the difference between R^2 and Q^2 values in Figure 42, 5-2-1 CNN model was selected as the final nonlinear WNV analysis model. The model had a low RMSE value of 1.78 and lowest difference (13 percent) between R^2 (0.75) and Q^2 (0.62)

values. The model had five input neurons each corresponding to one of the predictor variable chosen by the genetic algorithm, two hidden neurons, and one output neuron, which stored the predicted number of WNV infected dead birds for a zip code. The predictor variables selected for this model were distance to bogs (miles), distance to lakes (miles), daily maximum temperature (F), age of houses (years), and percentage of developed medium density land cover class. Figure 43 shows the lowest difference between the R^2 and Q^2 for the selected CNN model with 5 predictor variables and Figure 44 shows the model architecture.

Figure 41 Behavior of RMSE values with an additional increase of predictor variable and an additional increase of hidden neurons

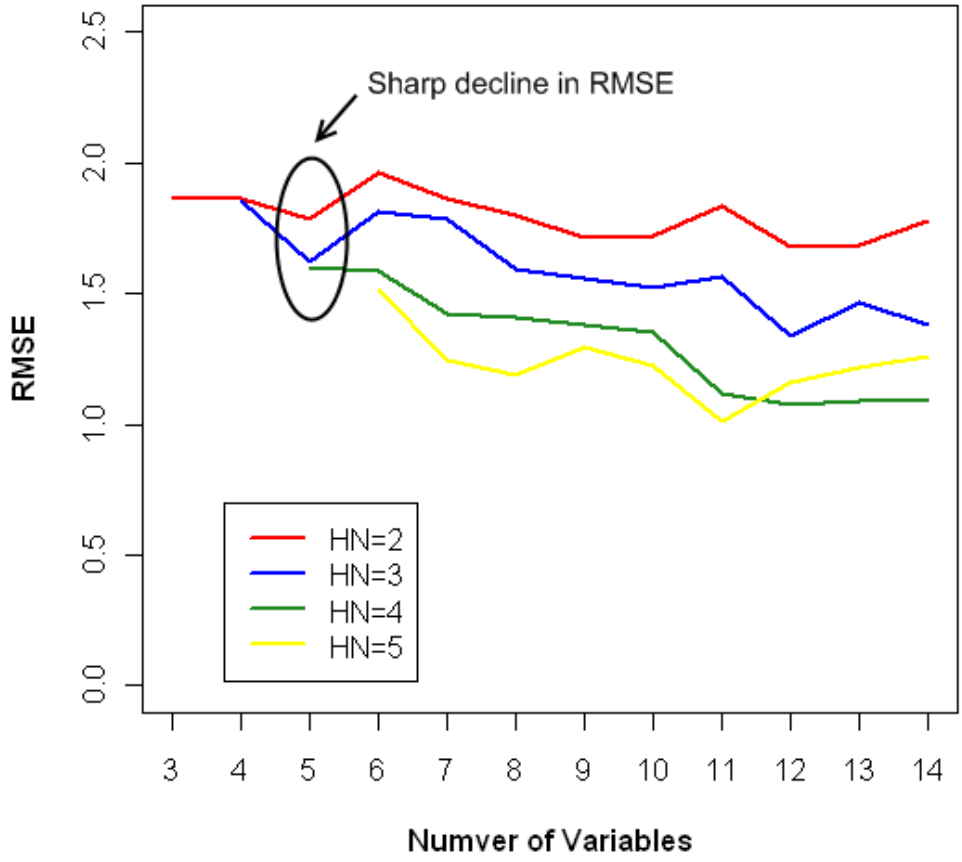


Figure 42 Difference between R^2 and Q^2 values with an additional increase of predictor variables and an additional increase of hidden neurons

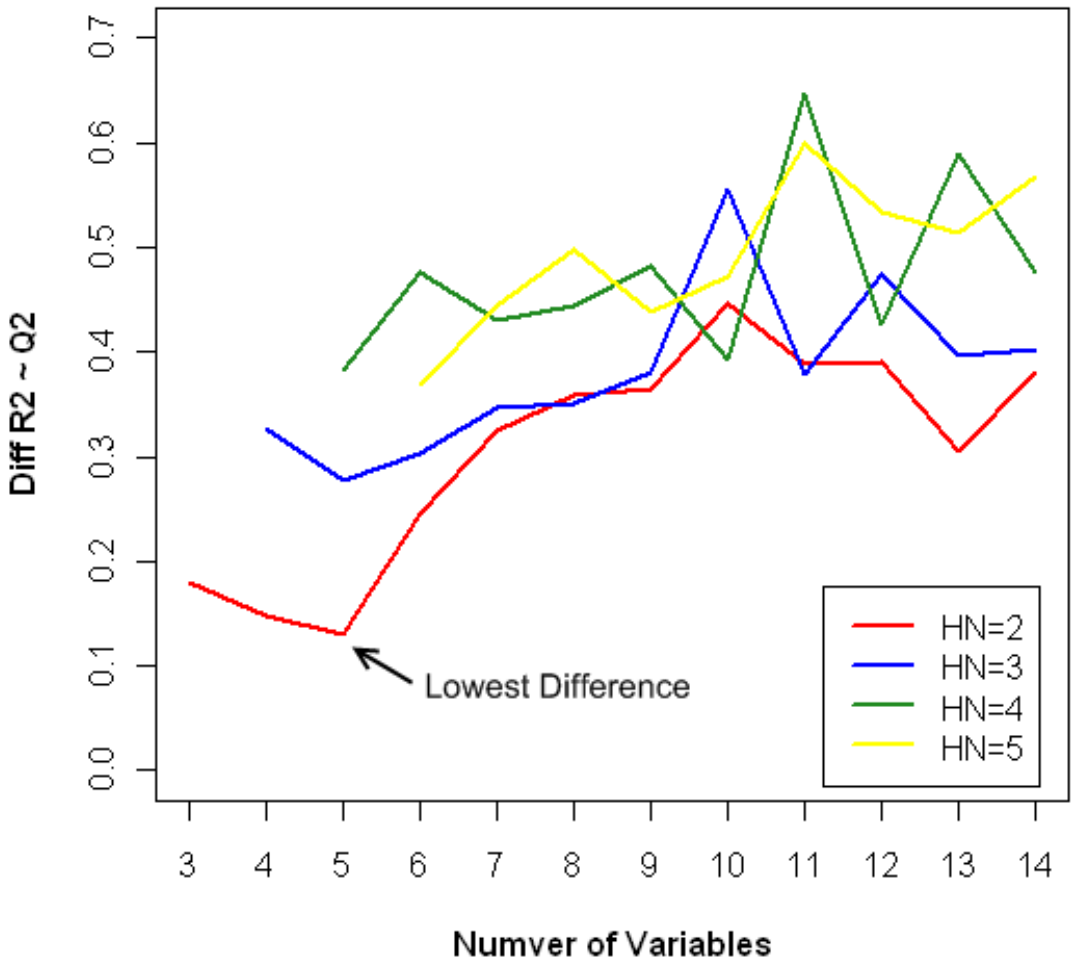


Figure 43 Trend of R^2 and Q^2 from CNN models with two hidden neurons and with additional increase of predictor variables

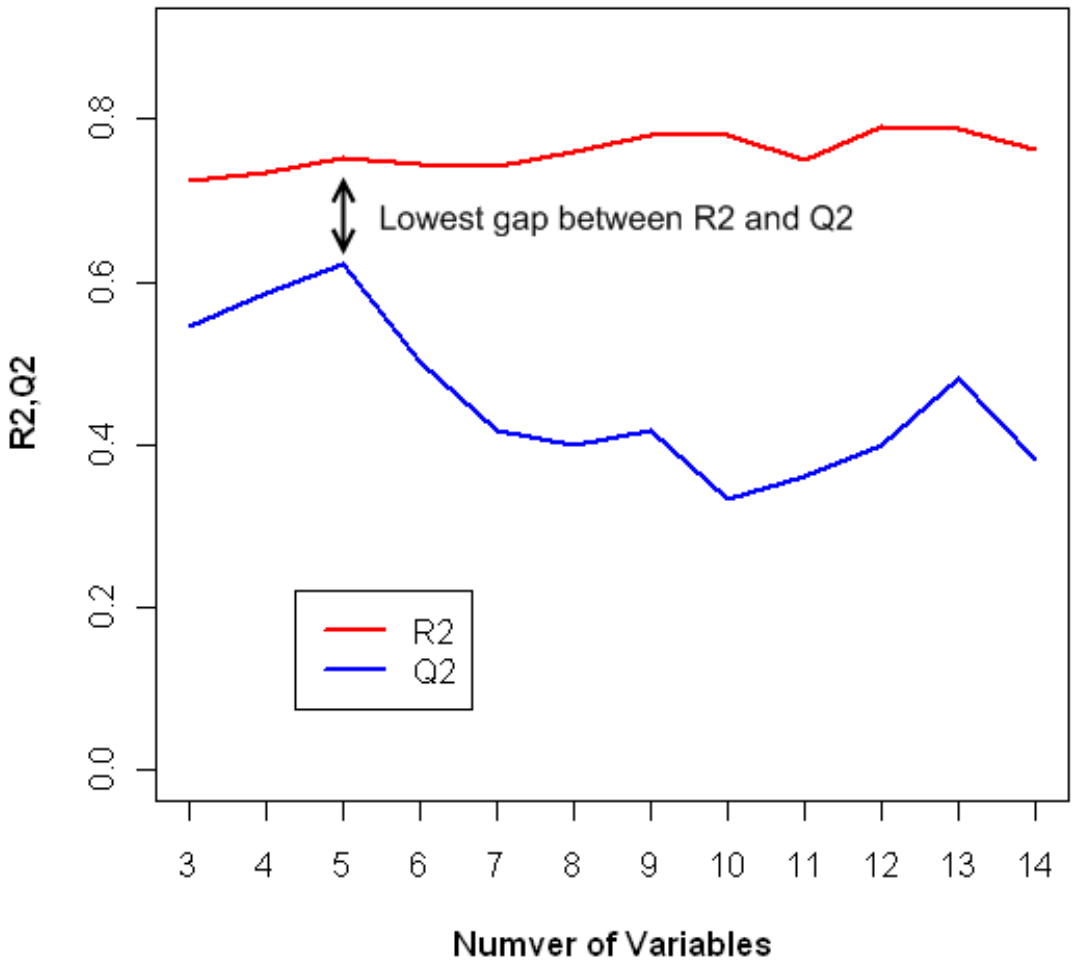
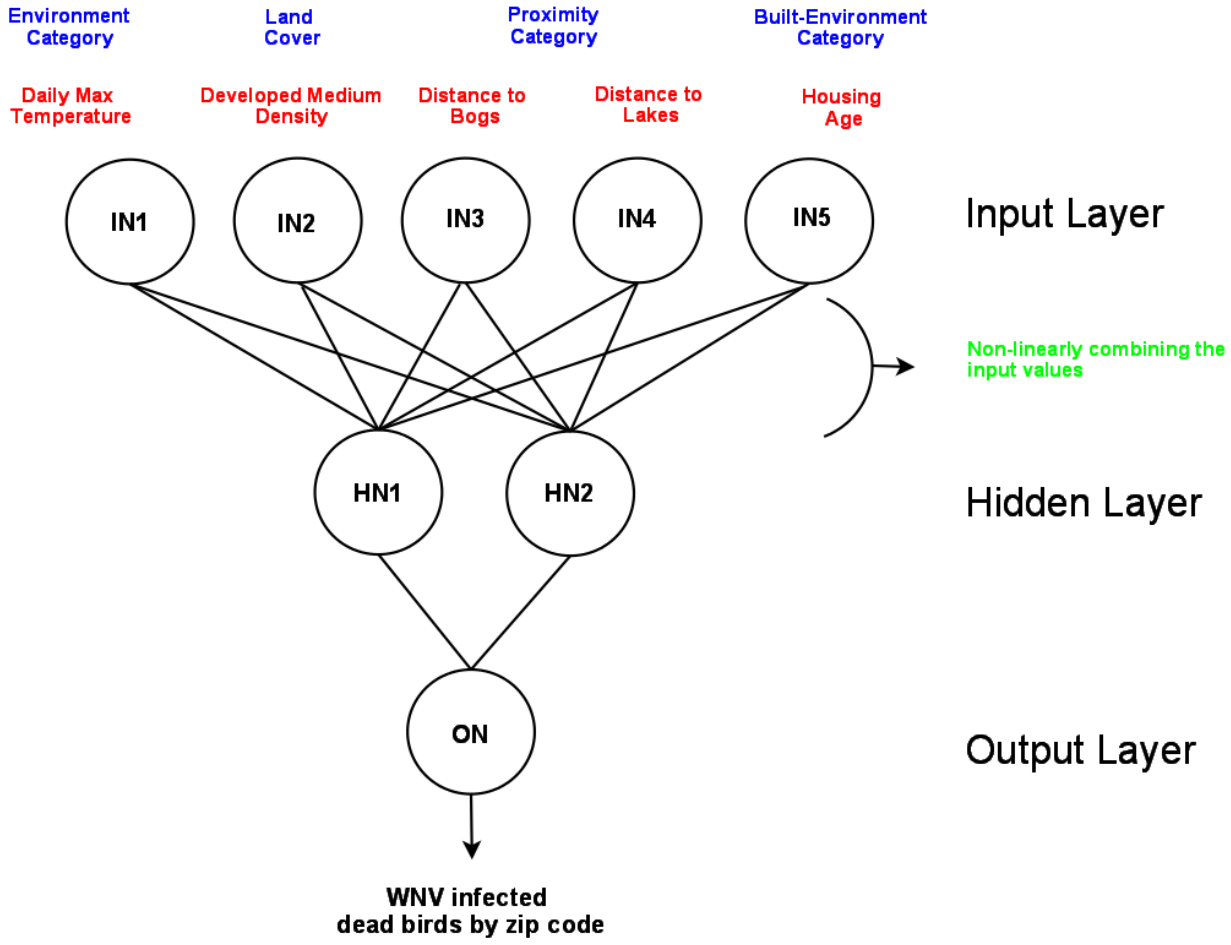


Figure 44 Structure of West Nile virus analysis Model using Feed-Forward Neural Network Algorithm



Note: IN = Input Neurons, HN = Hidden Neurons, ON = Output Neurons

To further assess the quality of the model, correlation between the observed and the predicted number of WNV infected dead birds were analyzed aspatially through histograms (Figure 45) and spatially by choropleth maps (Figure 46). In Figure 45, both the histograms show similar pattern with negatively skewed distribution i.e. there are

higher frequencies of zip codes with fewer counts of dead birds and vice versa. For both the observed and the predicted values, zip codes with 0 to 2 bird cases had highest frequency followed by zip codes with 3 to 5 cases and so on. Given the fact that, 74 out of 93 zip codes (80 percent) with 0 to 2 cases were correctly predicted indicated that the model performed well for lower values of the response variable. However the performance was relatively weaker at the tail of the histogram (i.e. for higher values of the response variable) indicating signs of under estimation.

Figure 45 Histograms of Observed and Predicted number of West Nile virus infected dead birds by zip codes in 2006

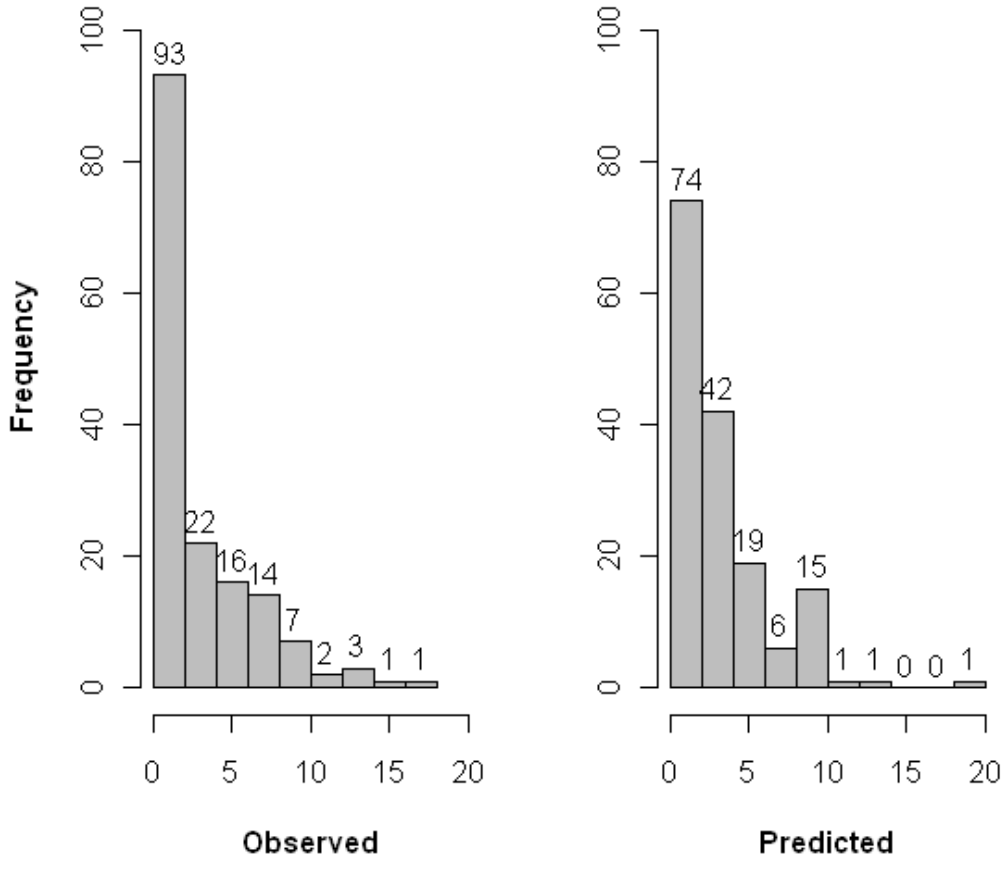
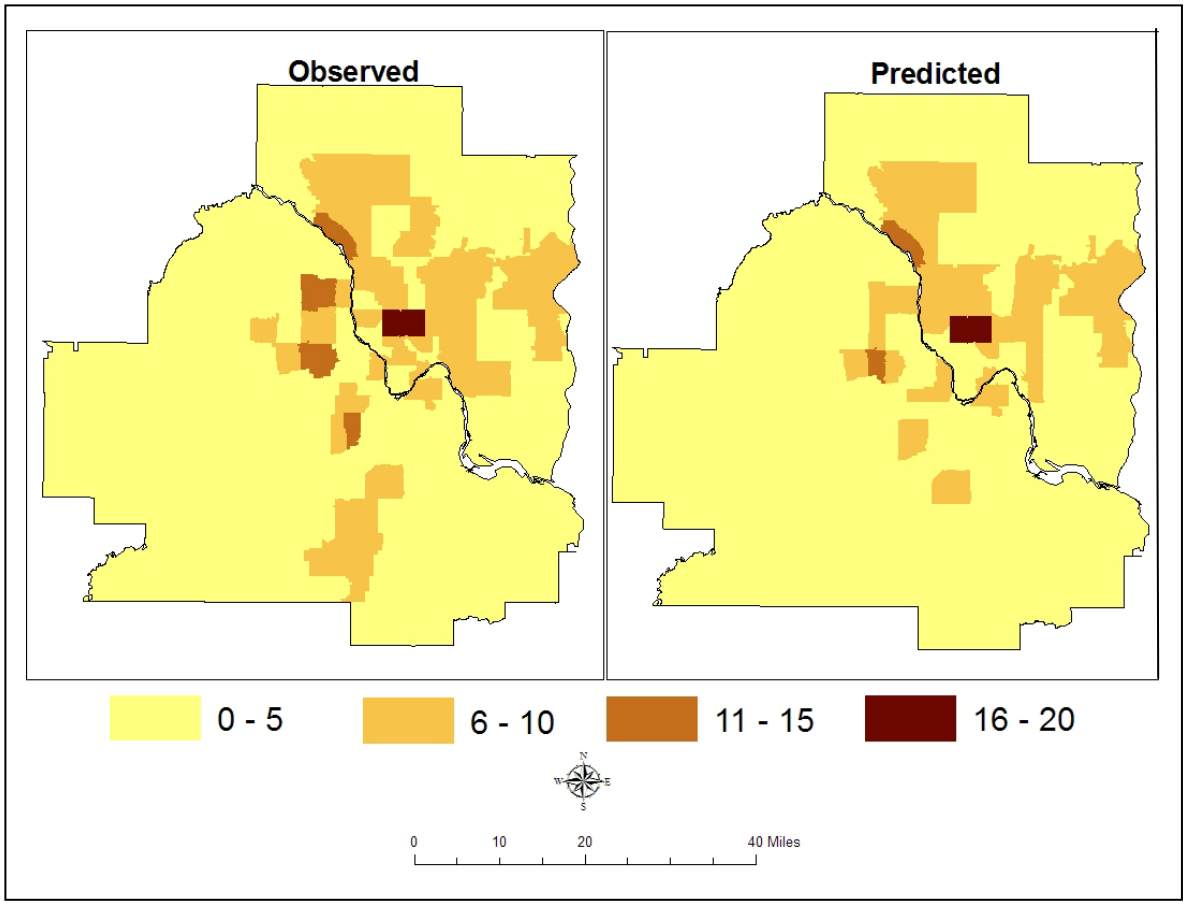


Figure 46 Spatial distribution of Observed and Predicted number of West Nile virus infected dead birds by zip codes in 2006



The choropleth maps of the observed and the predicted number of infected dead birds show similar spatial pattern with higher number of reports from the urban areas of the TCMA. This trend resonated with the hypothesis of urban preference of WNV transmission in the TCMA, discussed in the previous chapters. The prediction trend depicted in the histograms (Figure 45) was also observed on the choropleth maps, i.e., stronger for lower counts of dead birds and relatively weaker for higher counts of cases. Some of the zip codes with 11 to 15 cases were predicted with lower values of 5 to 10 cases. This resulted in higher frequencies of zip codes with 4 to 5 and 6 to 8 numbers of

cases (Figure 45) and thus the first class on the predicted map occupied more area (Figure 46).

In summary, the nonlinear 5-2-1 CNN model selected as the WNV analysis model had lower RMSE value (1.78), higher R^2 (0.75), lowest difference between R^2 and Q^2 values (13 percent), and similar aspatial (histograms) and spatial (choropleth maps) distribution of the observed and the predicted values of WNV infected dead birds.

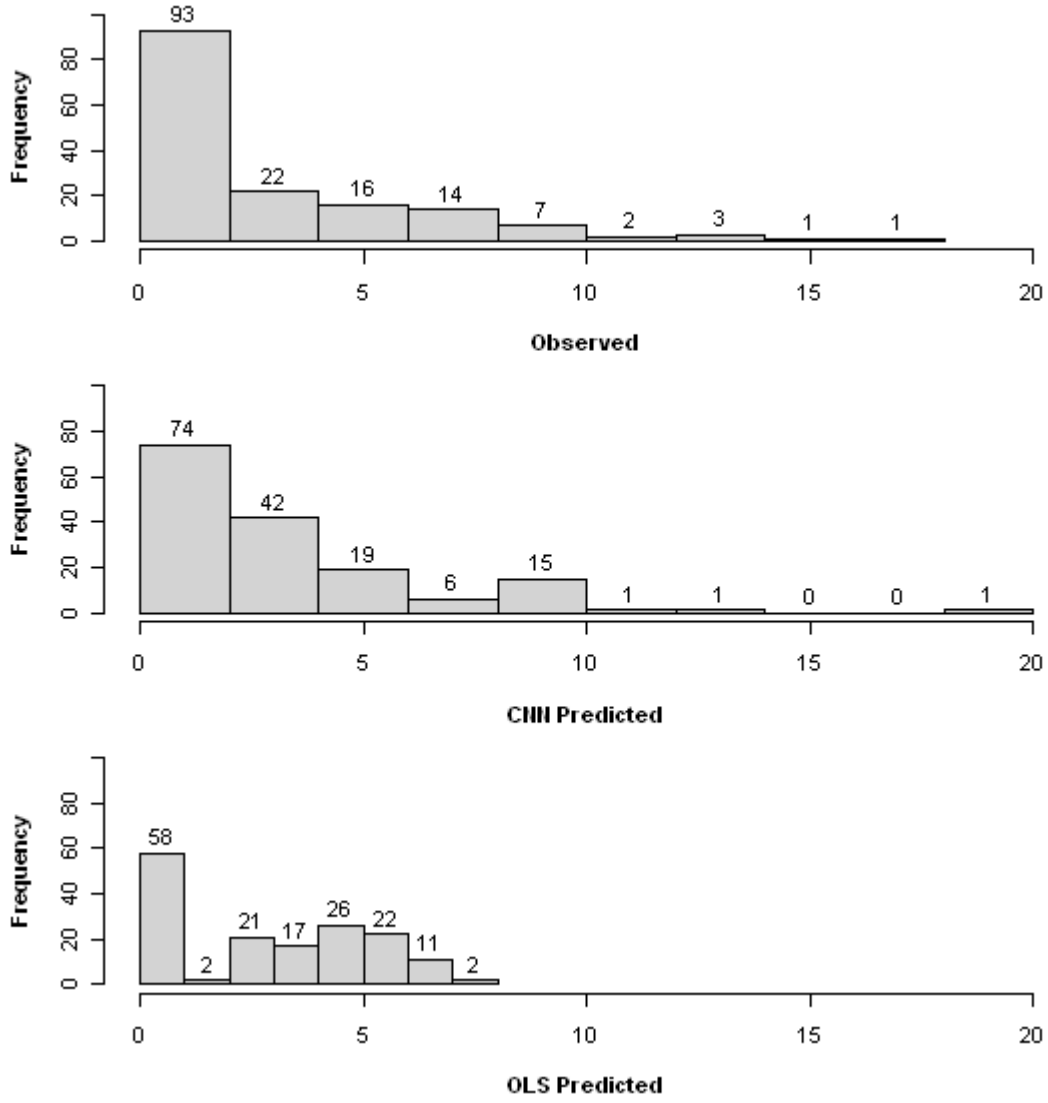
6.4.2 Comparison of CNN and OLS model

To compare the results of the selected WNV neural network model, an OLS model was build with same specifications. The OLS model had an *F*-statistic of 32.2 (for 5 and 153 degrees of freedom), which was much greater than the critical value of 1.96 ($\alpha = 0.05$). The model was thus statistically valid. The comparison of OLS and CNN models on the basis of RMSE, R^2, and Q^2 are shown in Table 27. Even though the difference between R^2 and Q^2 values was lowest in case of the OLS model, the CNN model performed better than the OLS model in terms of other indicators, such as lower RMSE and higher R^2 values. Moreover the internal cross-validation result (13 percent) was within the acceptable range of 10 to 15 percent.

Table 27 Comparison of OLS and CNN models with the best subset of predictors

Indicators	OLS	CNN
RMSE	2.56	1.78
R^2	0.51	0.75
Q^2	0.4	0.62
Difference	0.11	0.13

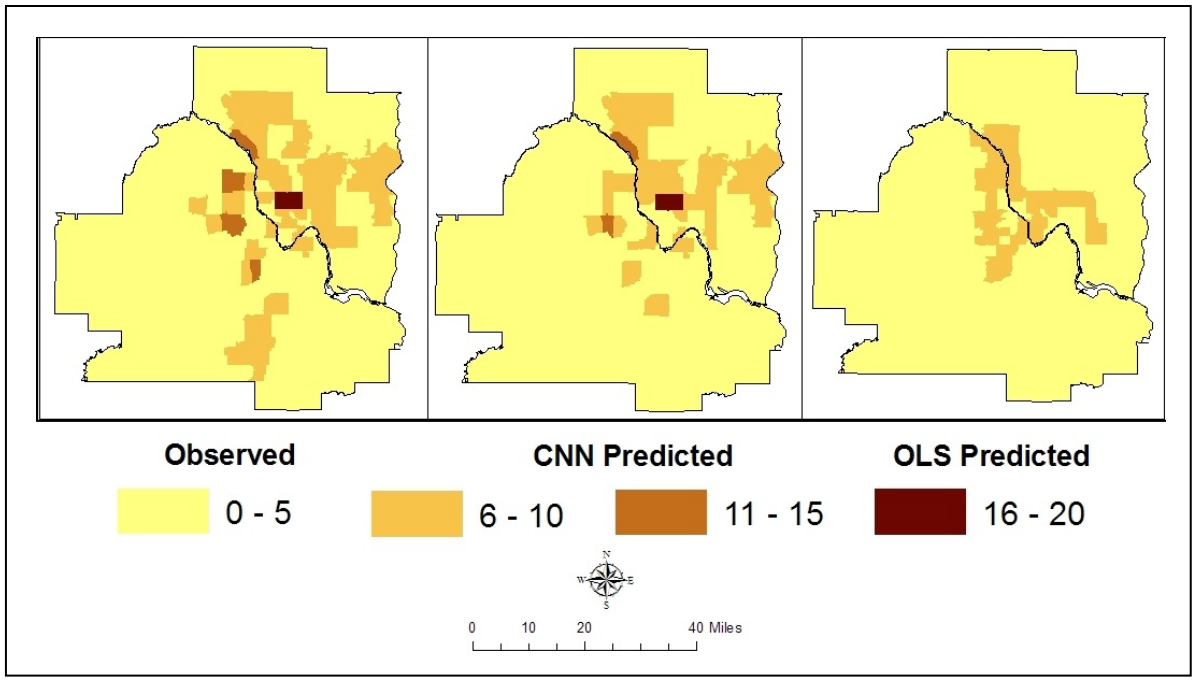
Figure 47 Histograms showing the comparison of observed, CNN predicted, and OLS predicted values of West Nile virus infected dead birds by zip codes in 2006



In Figure 47 the histogram of the observed data showed a negatively skewed distribution with higher frequencies of zip codes of no reports followed by 0 to 2 cases of dead bird reports and so on. As expected in a negatively skewed distribution, the

observed histogram had a thin and elongated tail indicating lower frequencies of higher dead bird counts. The CNN predicted histogram in overall showed similar pattern with strong correlation of the observed and predicted values for fewer numbers of infected dead birds and relatively weaker correlation for higher counts. On the contrary, the distribution of predicted values from the OLS model was significantly different from the observed data. The OLS predicted distribution was bimodal with highest frequency for 0 case count followed by 5 to 6 case counts. Out of 93 zip codes with 0 to 2 dead bird cases, only 58 (62 percent) were predicted correctly by the OLS model. On the other hand, the CNN model correctly predicted 74 of such zip codes. Another striking difference was that the OLS model underestimated the response variable. In the original dataset, the highest number of infected dead birds for a particular zip code was 17 cases. However the highest predicted value from the OLS model was only 9 cases and that of CNN model was as high as 19 cases. The under estimation by the OLS model is also reflected in the choropleth maps in Figure 48.

Figure 48 Spatial distribution of Observed, CNN Predicted, and OLS predicted number of West Nile virus infected dead birds by zip codes in 2006



Thus the comparative analysis of the OLS and CNN models on the basis of RMSE, R^2, Q^2, (Table 27), and aspatial (Figure 47) and spatial (Figure 48) distribution of the observed and the predicted values indicated that the 5-2-1 CNN model was superior to the OLS model in terms of quality and predictability. It also suggested that computational neural network models are better suited to capture nonlinear relationships between the predictors and response variable, such as involved in the dynamics of WNV infection.

6.4.3 Cross-Validation

The WNV predictive model was cross-validated using both the internal and external techniques described in the section 6.2.3. The LOO technique was used for internal cross-validation procedures and the results are reported in the section 6.4.1. For external cross-validation technique two external datasets were considered. These new datasets were observed number of WNV infected dead birds in the year 2003 and observed number of WNV infected dead birds in the year 2007. The data from the year 2003 was selected for retrospective prediction with 285 infected dead birds. A total of 60 infected bird cases from the year 2007 were chosen for prospective prediction. The use of these datasets will assess the predictive capability of the CNN model in two different scenarios of relatively higher (285) and lower (60) counts of infected dead birds. The descriptive statistics of the variables in the two datasets are shown in Table 28 and Table 29.

Table 28 Descriptive statistics of the 2003 dataset

Variables	Min	1st Quartile	Median	Mean	3rd Quartile	Max
Observed Y						
Bird_2003	0	0	1	2	3	10
Predictors						
Age of Houses (Years)	0	22	35	39	48	88
Distance to Bogs (Miles)	661.40	2834.70	5654.90	6995.50	8849.30	25297.80
Distance to Lakes (Miles)	60.59	301.48	440.46	524.30	646.11	1871.95
Dev, Medium Density (%)	0.62	4.41	17.03	18.80	30.06	64.28
Max Daily Temperature (F)	69.16	72.59	80.57	76.53	84.08	91.29

Table 29 Descriptive statistics of the 2007 dataset

Variables	Min	1st Quartile	Median	Mean	3rd Quartile	Max
Observed Y						
Bird_2007	0	0	0	0	1	6
Predictors						
Age of Houses (Years)	1	26	38	42	51	91
Distance to Bogs (Miles)	661.40	2834.70	5654.90	6995.50	8849.30	25297.80
Distance to Lakes (Miles)	60.59	301.48	440.46	524.30	646.11	1871.95
Dev, Medium Density (%)	0.62	4.41	17.03	18.80	30.06	64.28
Max Daily Temperature (F)	67.68	71.69	73.87	78.31	86.55	91.86

Neural Network models were built for both the datasets with the same 5-2-1 architecture of the WNV analysis model. The resultant RMSE, R^2, and observed verses predicted plots from the external datasets were analyzed to assess the predictive capability of the WNV model. Further, results were also compared with that of the OLS models with same specifications.

The RMSE and R^2 values of the 2003 dataset obtained from the CNN model were 1.01 and 0.65 respectively. Both the values were acceptable given that the R^2 was high and the RMSE value was close to 10 percent of the range of infected number of dead birds (0 to 10) in 2003. On the contrary, the RMSE value from the OLS model was 2.239, which was much higher than the CNN model. Similarly, the R^2 (0.488) obtained from the OLS was significantly lower than the neural network model (0.65). The histograms of observed verses predicted values are shown in Figure 49. Overall both the observed and the predicted histograms showed similar patterns of negatively skewed distribution with higher frequency of fewer dead birds counts and vice versa. In the original data, 92 zip codes reported no bird cases and 79 (86 percent) of such zip codes were predicted correctly by the CNN model. Further the CNN model was almost 100 percent accurate in predicting the dead bird counts for zip codes with number greater than eight. However the predictions were relatively weaker for the middle values especially for zip codes with 3 to 4 dead birds. In the observed distribution there were similar frequencies of zip codes with 3, 4, and 5 bird counts but the predicted histogram showed a bimodal distribution with 38 zip codes with 3 to 4 dead bird cases. In addition, Figure 50 also showed similar spatial pattern of observed verses predicted cases with generally higher number of infected dead birds reported at the zip codes around the Twin Cities of Minneapolis and Saint Paul.

Thus, except for the bimodal distribution of the predicted values, the overall similar trends of aspatial (histograms) and spatial (choropleth maps) distributions of the observed and the predicted values, higher R^2, lower RMSE, indicated that the CNN model had good predictive capabilities.

Figure 49 Histograms of Observed and Predicted number of West Nile virus infected dead birds by zip codes in 2003

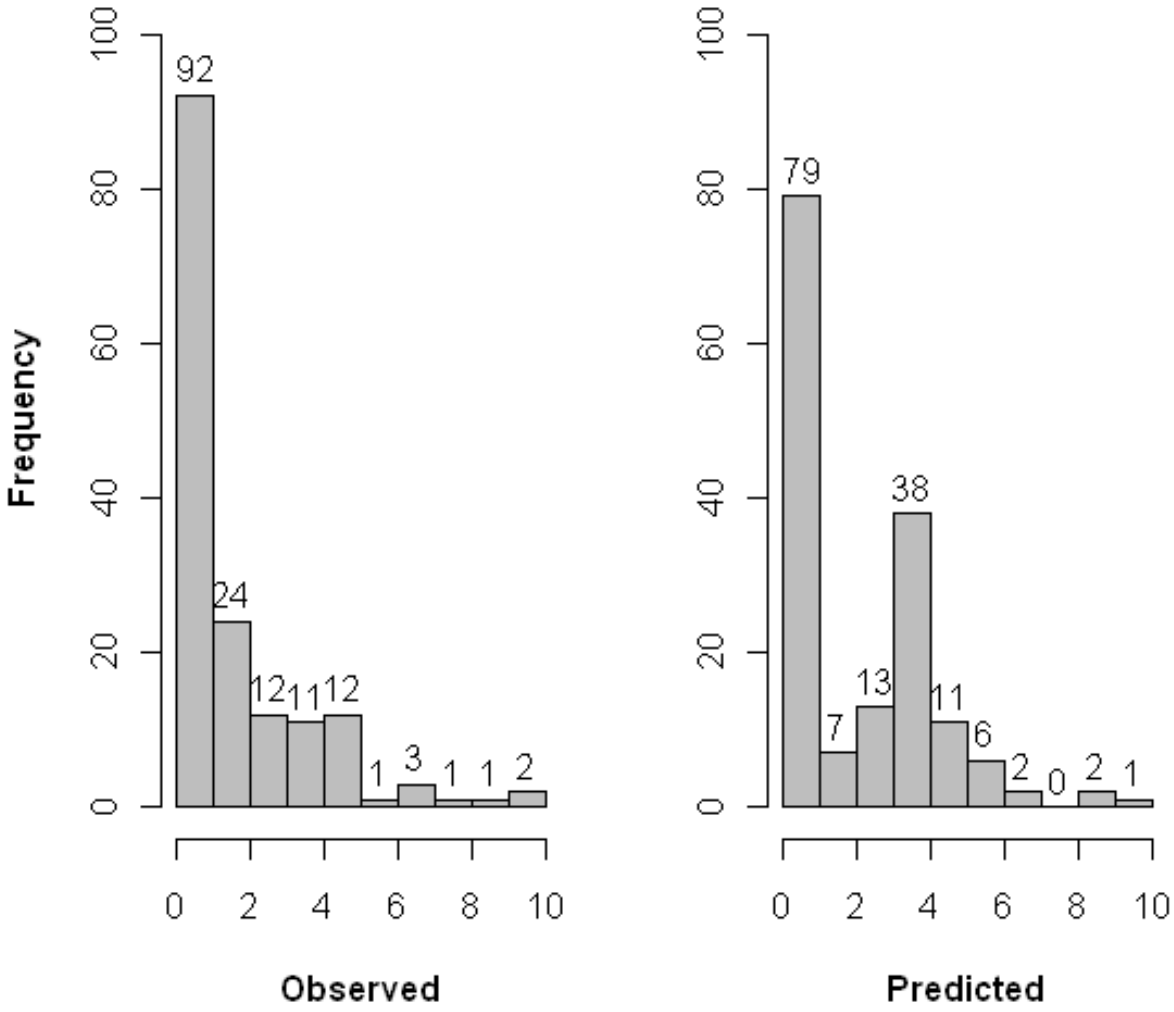
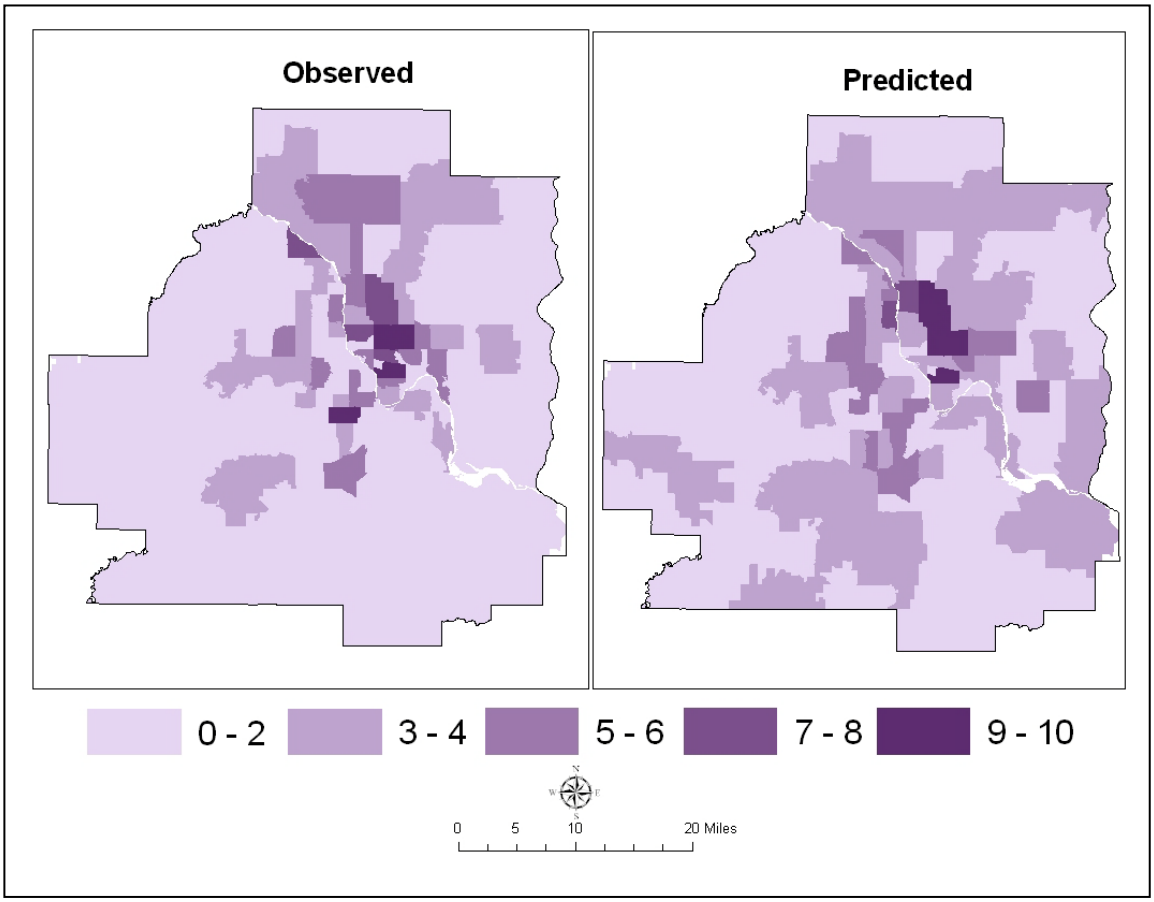


Figure 50 Spatial distribution of Observed and Predicted number of West Nile virus infected dead birds by zip codes in 2003



The RMSE and R^2 values of the 2007 dataset calculated from the OLS model were 2.33 and 0.47 respectively. The CNN model performed better than the OLS model with significantly lower RMSE value of 0.71 and higher R^2 value of 0.74. Both, aspatial distribution (Figure 51) and spatial distribution (Figure 52) of the observed and the predicted number of infected dead birds showed similar patterns. In Figure 51 both the histograms had similar distribution with highest frequency of zip codes with no cases, followed by zip codes with 1 case, and so on. Out of the 114 zip codes with no cases, 110 (96 percent) were predicted correctly. The strong correlation between the

observed and predicted values observed in the histograms was also reflected in the similar spatial distribution shown by the choropleth maps in Figure 52.

Figure 51 Histograms of Observed and Predicted number of West Nile virus infected dead birds by zip codes in 2007

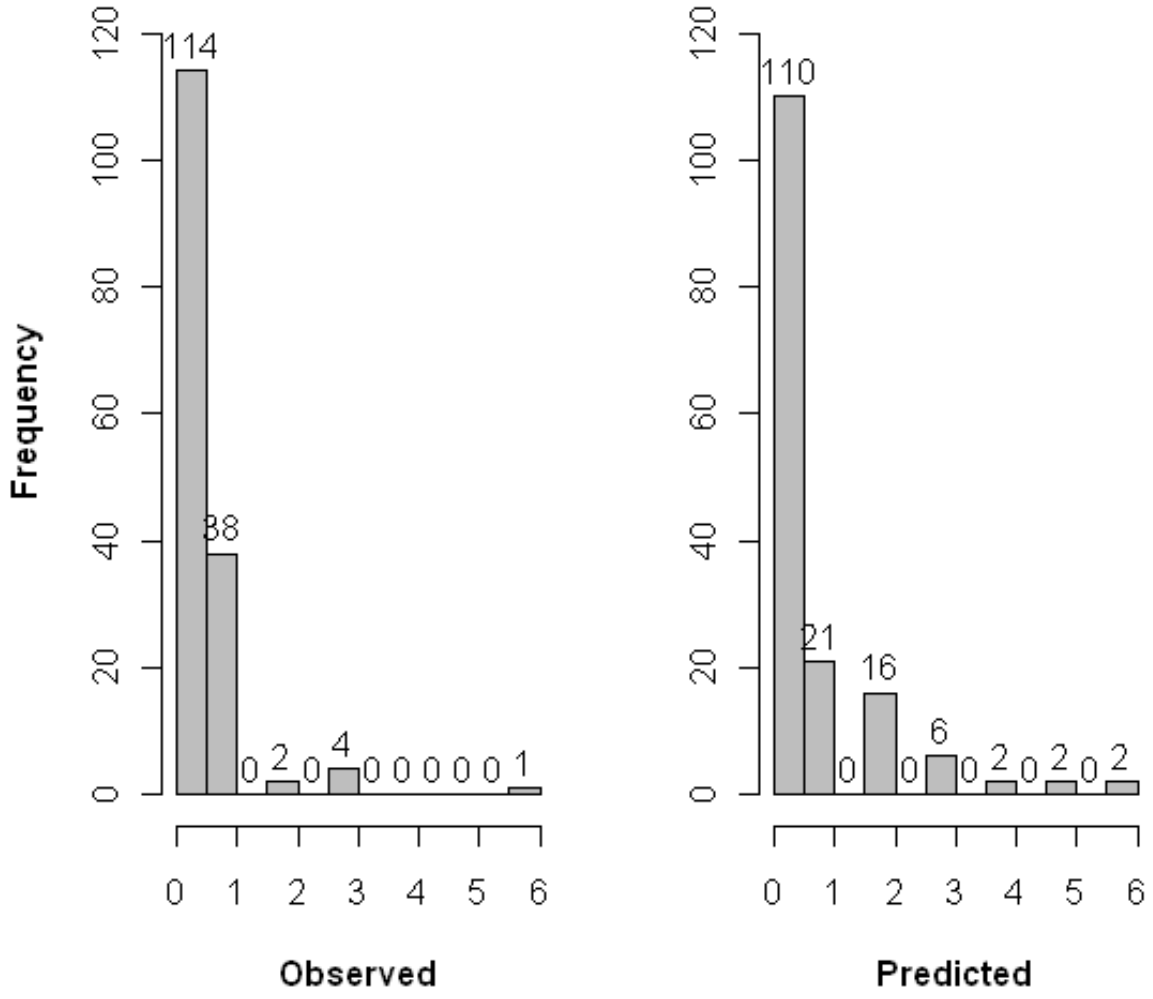
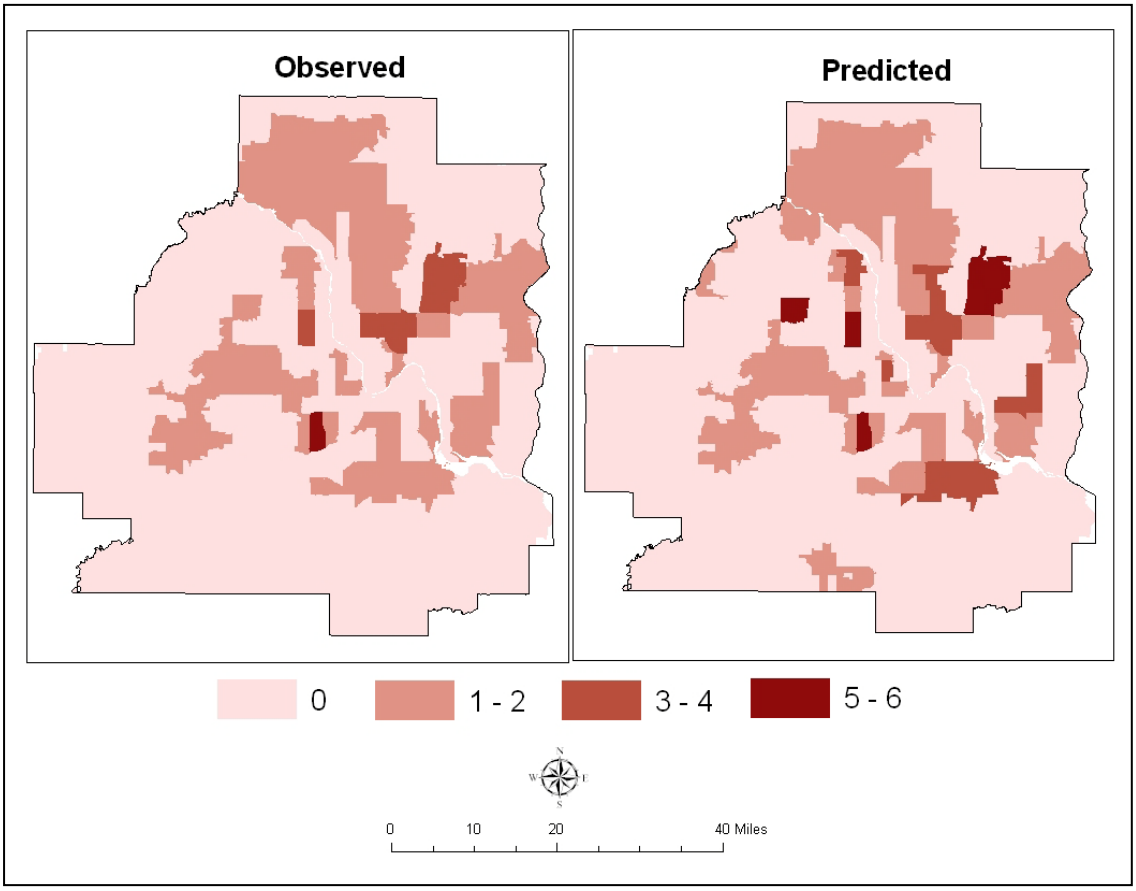


Figure 52 Spatial distribution of Observed and Predicted number of West Nile virus infected dead birds by zip codes in 2007



Thus from both the internal (LOO) and external (two external prediction datasets) cross-validation results, we can conclude that the nonlinear 5-2-1 CNN model selected as the WNV analysis model had good predictive capabilities. The RMSE and R^2 values obtained from the both the external prediction datasets were consistent. Finally, as expected, the CNN model had significantly lower RMSE values and higher R^2 values than the OLS models.

6.5 Conclusions

This chapter presented both linear and nonlinear neural network models to predict the number of WNV infected dead birds for the entire seven counties area of the TCMA. The specific goal of this work was to describe in detail the procedure involved in building a CNN model to predict the occurrences of WNV infected dead birds by incorporating the assumed nonlinearity between disease occurrence and hypothesized risk factors. The Genetic Algorithm was used to search the predictor space for the best subset of predictor variables which were then included in a CNN model. The final model selection was based on a combination of lower RMSE value, higher R^2, and lowest difference between R^2 and Q^2 values. This procedure resulted in a CNN model with 5-2-1 architecture as the best model for WNV analysis. The predictor variables included were distance to bogs, distance to lakes, daily maximum temperature, housing age, and percentage of developed medium density land cover class. Even though the OLS model with the same specification was statistically significant (*F*-statistic), the RMSE was significantly higher and R^2 lower than the values obtained from the CNN model. The observed versus predicted histograms and the choropleth maps suggested that the CNN model had much superior predictive power than the OLS model when modeling a complex health outcome, such as WNV. The external cross-validation results with new datasets also indicated that the selected CNN model had better predictive performance than the OLS models.

However neural network models have number of shortcomings. The most important criticism is its lack of interpretability or in other words its '*black box*' nature. As a result, the use of neural network algorithms is rare in Geography or any other social sciences. The limited use is restricted to only classification or prediction, rather than a technique used to *understand* or *explain* a phenomenon. The next chapter will attempt to address this shortcoming of CNN models by interpreting the neural network WNV model by two interpretation techniques, broad and detailed.

7. Chapter 7: Which are the contributing risk factors? How are they related? Interpreting West Nile virus computational neural network model

7.1 Background

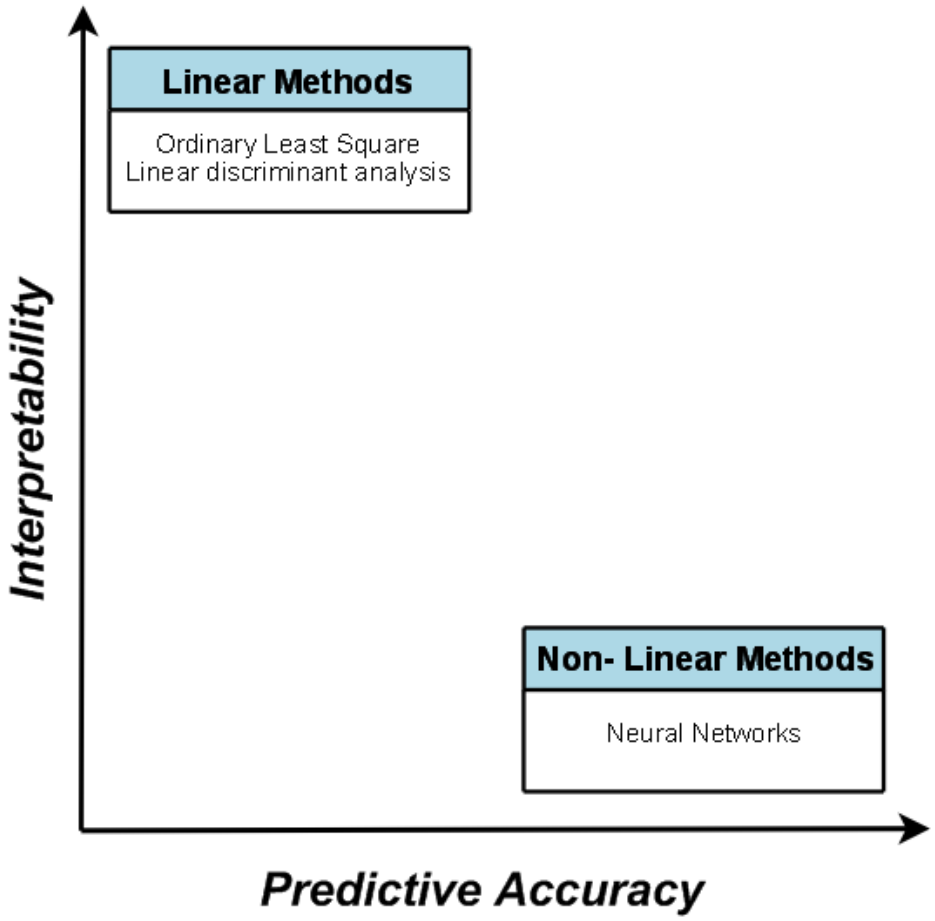
Computational neural networks (CNNs) are an important component of a modeler's toolbox for a number of reasons. First, neural network models have higher predictive ability than both linear (OLS) and traditional nonlinear (Logistic) techniques. Second, the higher predictive capability of CNNs arises from their flexibility (no *a priori* assumption of distribution, normality, independence, etc) and their ability to capture nonlinear relationships. The preceding chapter detailed the implementation of a neural network model capturing the complex dynamics of WNV transmission in the TCMA. Cross-validation techniques indicated that the model had superior goodness of fit and higher predictive performance than the OLS model with same model specifications.

Neural network models also have a number of shortcomings. The most important drawback is its *black box* nature or lack of interpretability. That is, one provides the input values and obtains an output value, but generally no information is provided regarding how those input values are associated with the output value. Even though the CNN model built in the previous chapter had higher predictive capabilities, it was obscure as to how the selected risk factors were associated with the occurrence of WNV infected dead birds in the TCMA. Along with accurate predictions, *interpretability* also plays an important role in modeling processes. Therefore

understanding or *explaining* the WNV analysis model, including the relationships between daily maximum temperature, age of houses, medium density land cover class, distance to bogs, and distance to lakes with the occurrence WNV infected dead birds, is crucial for several reasons. First, some measure of interpretability is needed to provide a sense of confidence regarding the soundness of the model. Second, detailed interpretation would also provide evidence to support the use of the model in future scenarios of WNV outbreaks. Third, we not only want to predict the future occurrence of WNV cases but also want to extract information about how the individual risk factors or combination of risk factors correlate to the disease occurrence. Fourth, the extracted relationships could be used for vector control policy recommendations. Fifth, the identified relationships between the risk factors and disease occurrence could shed light in the *causality* of the virus incidents and contribute to further research on WNV.

However, interpreting the encoded relationships in a CNN model is difficult. This often forces their use as a purely *predictive tool* rather than a technique to understand relationships. This is in contrast to linear models, which can be interpreted in a simple manner but have poor predictive ability. We observe that in many cases, the interpretability of a model is a trade-off with predictive accuracy, shown schematically in Figure 53. This is especially common in social sciences where understanding a phenomenon (e.g. land use change, disease occurrence, human-environment interactions, or crime) is often more important than prediction.

Figure 53 A Schematic diagram highlighting the trade-off between predictive accuracy and interpretability



The applications of CNN models in social sciences were limited to *only* prediction purposes, for example in physical geography (Browne et al. 2007) and development and planning (Blunden et al. 1998). In environmental and pollution research there were few attempts to extract meaning from neural network models (Clair and Ehrman 1996; Chang and Hsiao 2004). Most of these techniques involved

sensitivity analysis, which provided a broad interpretation of importance (contribution) of predictor variables. The main disadvantage of this method is that it provides a very general overview and is not capable of describing in detail the nature of the correlation (positive or negative) between a given predictor variable and a response variable. In cartography, a combination of CNN algorithms and multivariate analysis of variance (MANOVA) were used to investigate the cognitive processes of learning basic GIS functions and spatial concepts (Lloyd and Bunch 2003). This study was a classic example of a researcher's dilemma where a CNN model was used for its ability to capture nonlinear relationships with higher predictive performance however, due to its lack of interpretability, the encoded relationships in a CNN model were extracted using a linear MANOVA technique. This usage of contradictory modeling techniques, CNN (nonlinear) and MANOVA (linear) to compensate each other's shortcoming (interpretation vs. prediction) with the same underlying data surely indicates the need of efficient interpretation techniques for neural network models.

The majority of neural network applications in health studies were for accurate and consistent prediction of health outcomes, such as, colorectal and breast cancer (Burke et al. 1997), prostate cancer (Bostwick et al. 2005), schistosomiasis (Hammad et al. 1996), asthma (Hirsch et al. 1997), patient death in emergency room due to sepsis (Jaimes et al. 2005), and breast cancer screening (Ronco 1999). Attempts to interpret CNN models used in health related researches were limited (Kiang et al. 2006; Pan et al. 2008). In both these studies, sensitivity approaches were employed to rank the contributing predictor variables, but it did not provide any understanding of the correlation between the predictor variables and the response variable.

This chapter will address this important research need in the use of CNN models in social and health sciences by interpreting the WNN neural network model by two techniques, *broad* and *detailed*. The aim of a broad interpretation is to characterize how important a input variable (or risk factor) is to the predictive ability of the model (Guha 2005b). It is essentially a sensitivity analysis, which allows us to rank risk factors in the order of importance in explaining WNV. For example, this interpretation will indicate

which risk factor among distance to bogs and distance to lakes contributed most to the predicted incidence of WNV. However, as we have seen in several applications of CNNs, sensitivity approach do not allow us to understand exactly how a specific predictor affects the outcome of the model. Therefore the goal of a detailed interpretation is to extract the relationships between the predictor variables and the network output. A detailed interpretation should be able to indicate the directional association between a predictor and the response variable (Guha 2005c). This approach allows one to gauge the strength and direction of the effect of risk factors in explaining the occurrence of WNV. For instance, it will help to indicate whether a decrease in the proximity to lakes will increase the number of WNV cases. These broad and detailed approaches would therefore expand the role of CNN models in social and health sciences as both predictive and explanatory tool, thus alleviating the black box nature of the neural network methodology to some extent.

This chapter is organized as follows. The section 7.2 discusses the underlying details of the broad and detailed neural network interpretation techniques and section 7.3 recaps the characteristics of WNV neural network model explained in the previous chapter. The section 7.4 reports the broad and detail interpretation results and also compares them to that of the WNV OLS model. Section 7.5 provides a discussion of interpretation results, which are used in terms of vector control policy recommendations explained in section 7.6 and finally section 7.7 concludes the chapter.

7.2 Methodology

The following sections describe the methodology underlying the broad and detailed interpretation techniques for neural network models.

7.2.1 Broad Interpretation

In this study only 3-layer, fully connected, feed-forward computational neural networks were considered, where the input layer represented the predictors and the value of the output layer was the predicted value of the response variable.

The broad interpretation is essentially a sensitivity analysis of the neural network, which is viewed here as a measure of predictor importance. The algorithm to measure predictor importance adapted from the literature of computational chemistry (Guha 2005b) are as follows:

1. To start, a neural network model is trained and validated. The RMSE obtained from this model is defined as the base RMSE.
2. The first input predictor variable (say distance to bogs) is randomly scrambled, and then the neural network model is rebuilt and used to predict the number of WNV infected dead birds. Because the values of this predictor are shuffled, one would expect the correlation between distance to bogs and WNV occurrence in birds are obscure. As a result, the RMSE calculated from these new predictions should be larger than the base RMSE and the difference between these two RMSEs indicates the importance (contribution) of distance to bogs to the model's predictive ability. That is, if an input factor contributes significantly in the model's prediction, scrambling that factor will lead to a greater loss of predictive power (as measured by the RMSE value) than for an input factor that does not play such an important role in the model.
3. This procedure is then repeated for all the predictor variables present in the model.
4. Finally the predictor variables are ranked in order of their importance (difference between the base RMSE and RMSEs obtained from new predictions)

7.2.2 Detailed Interpretation

The detailed neural network interpretation methodology is analogues to the

procedure used for the interpretation of linear models using partial least squares (PLS) (Guha 2005c). Therefore I start with a short summary of the PLS technique.

7.2.2.1 Partial Least Squares

The predictors for a linear model are used to build a PLS model. The PLS model consists of latent variables (components) which are linear combinations of the original predictor variables. In the situation of no over fitting, the number of latent variables is equal to the number of input variables. The results of the PLS analysis are summarized by two tables. The first table reports the cumulative variances for each component. Typically the first few latent variables or components explain a large portion of the total variation (80% ~ 90%). As a result, the remaining components are ignored. The second table lists the X-weights for each component. These correspond to the linear combination of coefficients for each input variable in a given component. Therefore analysis of these weights allows one to understand the importance and correlation (direction) of a given input variable to the value predicted by that component.

7.2.2.2 Similarity of PLS and Detailed CNN interpretation technique

The detailed CNN interpretation method is based on the assumption that the hidden neurons are analogous to the latent variables or components in a PLS model. In addition, X-weights (linear combination of input variables) are also assumed similar to the connection weights (nonlinear combinations of input values) between the layers. Therefore by considering the weights connecting the input factors to a specific hidden neuron, we can then interpret how each predictor correlates to the output of that hidden layer neuron. Finally, by defining the contribution of each hidden layer neuron to the output value of the neural network, we can determine which hidden layer neurons are important and which ones can be ignored. The similarities between the two techniques are explained further in Table 30.

Table 30 Similarity of PLS and Detailed CNN interpretation technique

Partial Least Square		Detailed Interpretation Technique
Input Variables	→	Input Neurons
Latent Variables or Components	→	Hidden Layer Neurons
Rank Components based on variance explained	→	Rank hidden layer neurons based on SVC values (explained later)
X-weights	→	Connection Weights between the layers

7.2.2.3 Combining Weights

The detailed CNN interpretation is dependent upon how the weights and biases modify the input values as they pass through the layers on the network. We denote the weights between the input layer neuron *j* and the hidden layer neuron *i* as

$$w_{ij,}\ 1 \leq i \leq n_I \text{ and } 1 \leq j \leq n_H$$

where n_I is the number of input layer neurons (i.e. predictor variables) and n_H is the number of hidden layer neurons. Now let us consider the value of the first predictor for a given observation and how it passes through the layers.

As this value passes from the first input neuron to the first hidden layer neuron, w_{11} weight will be formed. The value from the first hidden layer neurons are then passed to the output neuron with the weight of w_1^H. Thus we can say, that, as the input value passes from the input layer to the hidden layer and then to the output layer, it is affected by a combined weight of $w_{11}w_1^H$. Similarly, for the same input value passing through the second hidden neuron and then to the output neuron, we can write the associated combined weight as $w_{12}w_2^H$. In general the combined weight between the

i^{th} input neuron and the j^{th} hidden layer neuron will be $w_{ij}w_j^H$ (Guha 2005c).

The absolute value and sign of $w_{ij}w_j^H$ is our main interest in terms of interpreting the correlation between the predictors and the response variable. The absolute value of the weights between the hidden layer neurons and the output neuron might be an indication of which hidden neuron is more effective in terms of contribution to the final output value. On the other hand, the sign of the weight indicates the trend of the output value. For example, if both the weights w_{11} and w_1^H are positive (or negative) we can expect that input values passing down that path will show a positive (or negative) correlation with the output value. If w_{11} and w_1^H are positive and negative respectively, one would expect that the net effect would be a negative correlation between the input values and output values.

Now let us see how the combined weights from the hidden layer neurons are associated with the network output. The output value of a CNN for a given set of input value is obtained via a sigmoidal transfer function. Thus we can write the output value, *O*, as

Equation 7.1 $$O = \frac{1}{1+\exp(-X)}$$

where *X* is the sum of weighted outputs from all the hidden layer neurons. If the output of each hidden layer neuron is denoted by x_j^H, $1 \leq j \leq n_H$, and the weight between each hidden layer neuron and the output neuron as w_j^H, $1 \leq j \leq n_H$, we can write *X* as,

$$X = \sum_{j=1}^{n_H} w_j^H x_j^H$$

Equation 7.1 can be rewritten as

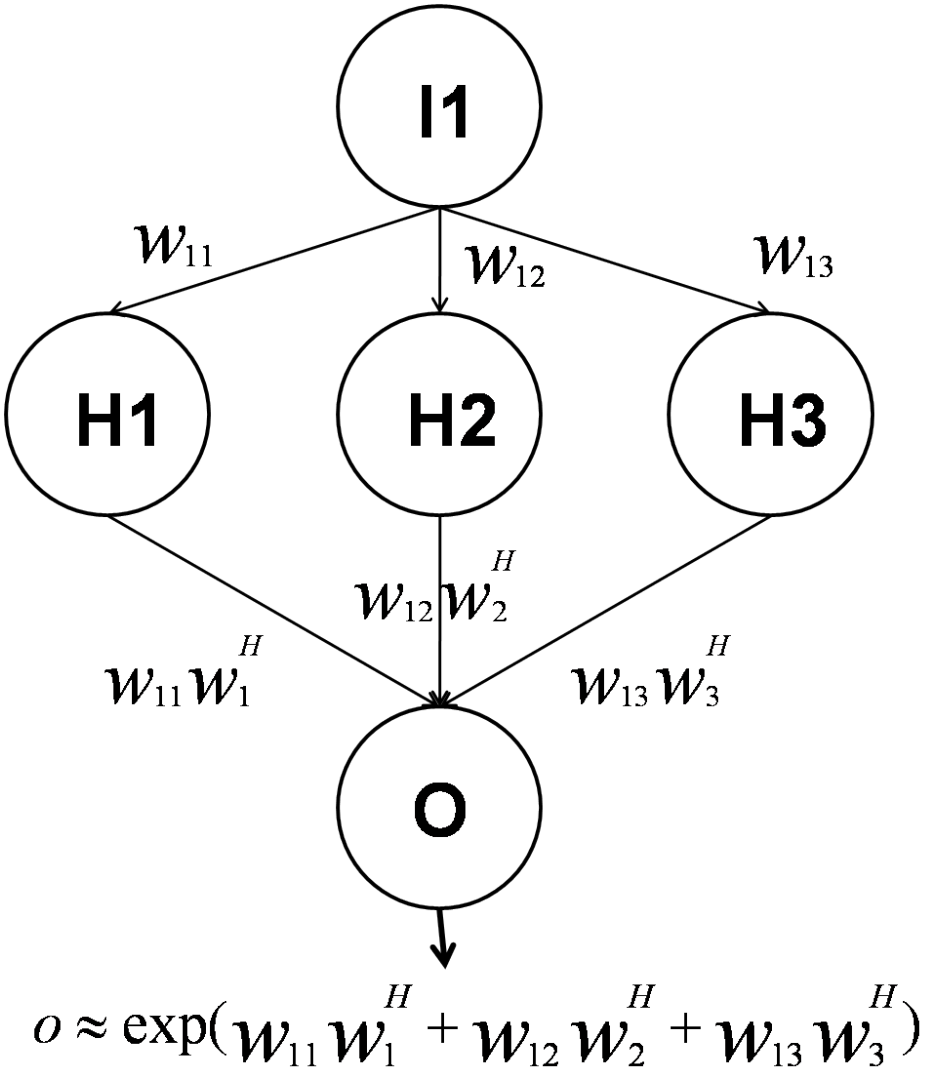
$$O = \frac{1}{1 + \exp\left(-\sum_{j=1}^{n_H} w_j^H x_j^H\right)}$$

$$\frac{1}{O} \sim \exp\left(-\sum_{j=1}^{n_H} w_j^H x_j^H\right)$$

Equation 7.2 $$O \sim \exp\left(w_1^H x_1^H + w_2^H x_2^H + \ldots\ldots\ldots + w_{n_H}^H x_{n_H}^H\right)$$

Equation 7.2 indicates that if a weight between a certain hidden layer neuron and the output neuron is large, then the output from that hidden neuron will dominate the sum. This allows us to rank the hidden layer neurons based on their contribution to the output value. Furthermore the signs of the weights indicate whether the hidden layer neurons will affect the output value positively or negatively (Guha 2005c). Figure 54 further explains schematically the accumulation of weights as the value from one input neuron flows down the network with three hidden layer neurons and one output neuron.

Figure 54 Schematic Diagram of Combined Weights flowing down the layers in a hypothetical 1-3-1 CNN model for a given observation



Note: I1 = Input Neuron (or predictor), H1, H2, H3 = Hidden Layer Neurons, O = Output Neuron

7.2.2.4 Interpreting Combined Weights

In this section we will consider two possible ways to use the weights to interpret the relationships between the predictor variables and the response variable. From the

preceding discussion we can represent the weights from a hypothetical 4-3-1 CNN model in a matrix form as shown in Table 31. Here, I1, I2, I3, I4 represents the four input neurons (predictors), $w_{ij,}$ denotes the connection weight between the i^{th} input neuron and the j^{th} hidden neuron, and w_j^H represents the weight between the j^{th} hidden neuron and the output neuron. For this example i ranges from 1 to 4 and j ranges from 1 to 3.

Table 31 Tabular representation of combined weights for a hypothetical 4-3-1 CNN model

	Hidden Neurons		
	1	2	3
I1	$w_{11}w_1^H$	$w_{12}w_2^H$	$w_{13}w_3^H$
I2	$w_{21}w_1^H$	$w_{22}w_2^H$	$w_{23}w_3^H$
I3	$w_{31}w_1^H$	$w_{32}w_2^H$	$w_{33}w_3^H$
I4	$w_{41}w_1^H$	$w_{42}w_2^H$	$w_{43}w_3^H$

The first step in interpreting the weight matrix is to decide the order of the hidden layer neurons in terms of their contributions to the output value. To do so I have followed the ranking technique explained by Guha *et al*. (2005c). The authors calculated the "squared contribution value" (SVC) for each hidden layer based on the combined connection weights between the input layer and the hidden layer and then between the hidden layer and the output layer. Individual SVC value ranges from 0 to 1 and they sum to 1 for all the hidden layer neurons. SCV values are functionally analogous (but not mathematically equivalent) to the cumulative variance explained by the latent variables or the components in a PLS technique. Therefore, the SCV values clearly indicate the contributions of each hidden neuron and allow us to possibly ignore

hidden neurons that have very small values of SCV.

For this hypothetical 4-3-1 CNN model, let us assume that the order of importance of the hidden layer neurons is given by H1 > H2 > H3. Thus, the first hidden neuron is the main contributor to the output value. Next we consider the values in the first column (H1) to find out which input neuron is associated with the highest weight. If the value in a given row is higher than the others it implies that the corresponding input neuron (predictor) has contributed more to the hidden layer neurons. Since we have already ordered the hidden neurons, this means that we can identify the contribution of each input neuron to the output value. Furthermore the sign of the weights will indicate whether high values of that input neuron correspond to low or high values of the output value.

7.3 Description of WNV analysis model

The nonlinear neural network WNV model built for the year 2006 had 5-2-1 architecture, i.e., five input (predictors) layer neurons, two hidden layer neurons, and one output layer neuron. The five predictor variables each corresponding to one input neuron were distance to lakes (miles), distance to bogs (miles), maximum daily temperature (F), age of houses (years), and percentage of medium density land cover class. The descriptive summary of the risk factors are reported in Table 32. The hidden neurons were responsible for nonlinearly combining the input values obtained from the input layer. The combined weights and biases from the hidden layer were then transferred to the output layer. Finally the output neuron predicted the network output of WNV infected dead birds. The model was trained and validated by both the internal (Leave-one-out and LOO) and the external (new datasets for the year 2003 and 2007) cross-validation techniques. The RMSE was 1.78 and R^2 was 0.75. The Q^2 value obtained from the LOO cross-validation technique was 0.62.

Table 32 Descriptive summary of variables in the WNV neural network model

Variables	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Unit
Response Variable							
WNV Infected Dead Birds	0	0	2	3	5	17	count
Predictor Variables							
Daily Max Temperature	75.39	80.31	82.93	79.52	84.48	98	F
Developed, Medium Density	0	2.071	12.52	12.77	20.09	39.11	%
Age of houses	0	25	37	41.21	52	90	years
Bog (D)	661.4	2835	5655	6996	8849	25300	miles
Lake (D)	60.59	301.5	440.5	524.3	646.1	1872	miles

7.4 Interpretation Results

This section reports the results from both the broad and the detailed interpretation of WNV neural network model. The broad interpretation approach measures the importance of predictor variables included in the model in terms of their contribution to the response variable and the detailed interpretation goes one step further and teases out the encoded relationships between the predictors and the response variable.

7.4.1 Broad Interpretation Results

We first consider the results from the WNV OLS model. The statistics of the linear regression model are summarized in Table 33. The R^2 value was 0.51, and the F-statistic was 32.5 (for 5 and 153 degrees of freedom) which was much greater than the critical value of 1.96 ($\propto = 0.05$). The model was thus statistically significant.

Table 33 Summary of the linear regression model developed for the WNV model

Variables	Estimate	Std. Error	*t*	*P*	Sig
(Intercept)	-1.332	0.068	-1.993	0.048	*
Distance to Lakes	0.0004	0.00004	0.914	0.362	
Distance to Bogs	-0.094	0.00005	-1.998	0.041	*
Age of Houses	0.003	0.0012	0.11	0.913	
Dev, Medium Density	0.098	0.0018	5.093	0.000	***
Daily Max Temperature	0.429	0.0006	7.139	0.000	***

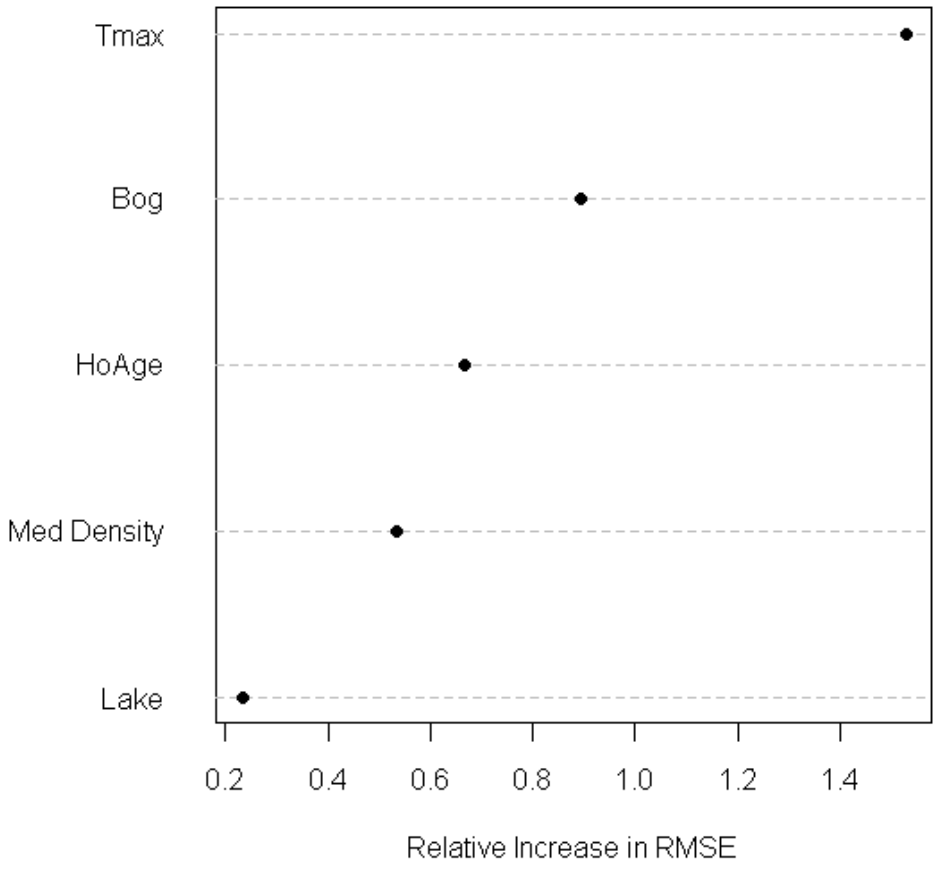
Based on the absolute values of parameter estimates (second column in Table 33), the variables contributing to the predicted number of WNV infected dead birds in decreasing order were maximum daily temperature, percentage of developed medium density land cover class, distance to bogs, housing age, and distance to lakes. However, if we only consider the parameter estimates which were statistically significant at $p < 0.05$, three risk factors were important. They were maximum daily temperature, land cover class, and distance to bogs. Temperature and land cover belonged to the environmental category and distance to bogs to the proximity category of hypothesized risk factors of WNV (see Chapter 2 for categorization of risk factors).

The next step was to measure the predictor importance of WNV CNN model with the same model specification. The increase in RMSE values for the predictors in the CNN model are reported in Table 34. The fourth column in the table represented the increase in RMSE due to the scrambling of the corresponding predictor variable, over the base RMSE of 1.78. It was evident that scrambling some descriptors led to larger increase, whereas others led to negligible increases in the RMSE. The information contained in Table 34 is more easily seen in the predictor importance plots shown in Figure 55. This figure plots the increase in RMSE for each input factor in decreasing order.

Table 34 Increase in RMSE due to scrambling of individual predictor variables. The CNN architecture is 5-2-1 with a base RMSE of 1.78

Scrambled Predictors	Description	RMSE	Difference
Tmax	Daily Max Temperature	3.308	1.528
Bog	Distance to Bogs	2.673	0.893
HoAge	Housing Age	2.446	0.666
Med Density	Dev, Medium Density	2.315	0.535
Lake	Distance to Lake	2.012	0.232

Figure 55 Importance Plot for the 5-2-1 West Nile virus CNN model



Considering Table 34 and Figure 55, we could say that the most important risk factor contributing to the occurrence of WNV infection in birds in the TCMA was maximum daily temperature. This risk factor represents an important environmental determinant associated to the occurrence of WNV incidences and is heavily documented in the literature (Theophilides et al. 2003; O'Leary et al. 2004; Ruiz et al. 2004; Hayes et al. 2005; Gibbs et al. 2006; Savage et al. 2006; Theophilides et al. 2006; Vaidyanathan and Scott 2006; Adlouni et al. 2007; Bolling et al. 2007; Gleiser et al. 2007; Landesman et al. 2007; Zou et al. 2007; Bouden et al. 2008; DeGroote et al. 2008; Ozdenerol et al. 2008). The second most important risk factor was distance to bogs. The Metropolitan Mosquito Control District (MMCD) defined bogs as one of the eight wetland types with higher potential for mosquito breeding. Previous studies also indicated that areas around wetland features namely bogs and swamps were at higher risk for the occurrence of WNV incidences (Rappole et al. 2000; Cooke, Katarzyna et al. 2006; Ezenwa et al. 2007; Leblond et al. 2007; DeGroote et al. 2008). In Figure 55 the significant gap between these two risk factors along the X-axis (relative increase in RMSE) qualitatively indicated that maximum daily temperature was contributing much more to the predicted output than the distance to bogs.

The next two important factors were housing age and percentage of developed medium density land cover, both of which belonged to the built-environment category of hypothesized risk factors. The land cover class is defined as areas with mixture of constructed materials and vegetation, impervious surface accounting for 20 – 49 percent of total cover, and is mostly occupied by single-family housing with typical grassy backyards and sometimes swimming pools (See Chapter 2 for details). The selection of these two risk factors and their level of importance was in tune with the urban-centric nature of WNV transmission in the Midwestern United States (Ruiz et al. 2004; Ruiz et al. 2007). The insignificant separation between these two variables along the X-axis also indicated that these two factors were probably playing similar roles in the predictive ability of the CNN model. The last risk factor in the order of importance was distance to lakes.

When compared to the predictor contributions from the OLS model, maximum daily temperature was ranked highest by both the models with significant separation (parameter estimates of the OLS model and relative difference in RMSE for the CNN model) from the second most important factor. Both the models also ranked distance to lakes as the least important factor. However the rankings of the other risk factors were different. For the CNN model, the second most important predictor was distance to bogs followed by age of houses and medium density land cover. On the contrary, the second most important predictor in the OLS model was medium density land cover class followed by distance to bogs and age of houses. The CNN model was able to separate the contribution of environmental variables (temperature and bogs) and built-environment variables (medium density and age of houses). However, this was lacking from the results of the OLS model. Such differences in results might be due to the fact that the nonlinear CNN model was a better fit and therefore was able to find distinct and good correlations between the predictors and the response variable than the linear model.

7.4.2 Detailed Interpretation Results

Similar to the broad interpretation, we first present a discussion of the OLS model. Based on the signs of the parameter estimates (second column in Table 33), we can understand the relationships between the predictor variables and the response value. Among the variables which were statistically significant at p-value < 0.05, maximum daily temperature and medium density land cover were positively related to the predicted output. In other words, higher values of maximum temperature could lead to higher numbers of infected dead birds and therefore amplify the rate of WNV transmission. Similarly areas with developed medium density land cover in the TCMA will have higher risk for WNV activities. Areas in and around older houses were also at higher risk but this relationship was not statistically significant at the specified p-value. The distance to bogs was negatively related to the occurrence of WNV infected dead

birds, i.e. areas near bogs or closer proximity to bogs would increase the risk of WNV infection. The last input factor, distance to lakes was positively related to the number of dead birds and was also not statistically significant. Here the positive correlation is interesting and needs further investigation because several studies indicated that closer proximity to hydrologic features including lakes, wetlands, and estuaries increase the risk of WNV infection in birds, mosquitoes, and humans (Rappole et al. 2000; Cooke, Katarzyna et al. 2006; Ezenwa et al. 2007; Leblond et al. 2007; DeGroote et al. 2008). The following paragraphs compare these results with those obtained from the detailed interpretation of the WNV CNN model.

The combined weight matrix obtained from the detailed interpretation of WNV CNN model is shown in Table 35. The columns corresponded to the hidden neurons and were ordered by the squared contribution value (SCV) shown in the last row of the table.

Table 35 The combined weight matrix for the 5-2-1 West Nile virus model

	Hidden Neurons	
	1	**2**
Distance to Lakes	-197.34	78.02
Distance to Bogs	-1000.95	-857.77
Age of Houses	732.54	569.11
Dev, Medium Density	380.18	282.27
Daily Max Temperature	1433.95	902.86
SCV	**0.78**	**0.22**

The SCV values in the Table 35 indicated that the first hidden neuron played an important role in explaining the model output of predicted number of infected dead birds. The second hidden neuron played a lesser role. The use of the SCV values in

choosing the influential hidden neurons was analogues to the use of percentage of variance explained by the latent variables in a PLS approach (Table 30). Considering the absolute values of weights in the second column of Table 35, we observed that the most weighted predictors were maximum daily temperature, distance to bogs, and housing age. Maximum daily temperature and housing age had large positive weights, indicating strong positive correlation between these risk factors and disease occurrence in birds. On the other hand, distance to bogs was associated with negative weights i.e., areas near bogs or closer proximity to bogs increased the risk of WNV infection. The other predictor variables, developed medium density land cover and distance to lakes had relatively smaller absolute values of weights with positive and negative signs respectively.

When we considered the second column (second hidden layer neuron), the order of predictor importance based on the absolute values of combined weights were same as that of the first hidden neuron. Except for distance to lakes, all the other predictor variables had same signs (correlation) as before, thus confirming the relationships between these predictor variables and the number of infected dead birds. For the first hidden neuron, the distance to lakes had negative correlation to the WNV infection in birds, i.e. with closer proximity to lakes, the risk of infection among birds increased. However the relationship became positive under the second hidden neuron. Given the fact that the first neuron had significantly higher contribution to the predicted output than the second hidden neuron, and that the absolute values of weights corresponding to distance to lakes was double in first hidden neuron (197.34) than the second hidden neuron (78.02), we can confidently say that the correlation between the proximity to lakes and the number of WNV infected dead birds was negative. Several previous investigations of WNV and its risk factors showed similar results (Rappole et al. 2000; Cooke, Katarzyna et al. 2006; Ezenwa et al. 2007; Leblond et al. 2007; DeGroote et al. 2008).

The relationships that emerged between the predictor variables and the response variable is also shown in a schematic diagram in Figure 56. We can also visualize the

positive correlation between maximum daily temperature and number of infected dead birds in Figure 57, negative correlation with distance to bogs in Figure 58, positive relationship with housing age and developed medium density land cover in Figure 59 and Figure 60 respectively, and lastly negative correlation between distance to lakes and WNV infection in birds in Figure 56. In all of the above mentioned figures, the maps on the left represent each of the predictor variables and the map on the right shows the number of WNV infected dead birds predicted by the 5-2-1 CNN model.

Figure 56 A schematic diagram showing the relationships between the risk factors and occurrence of West Nile infected dead birds obtained from the detailed interpretation of the CNN model

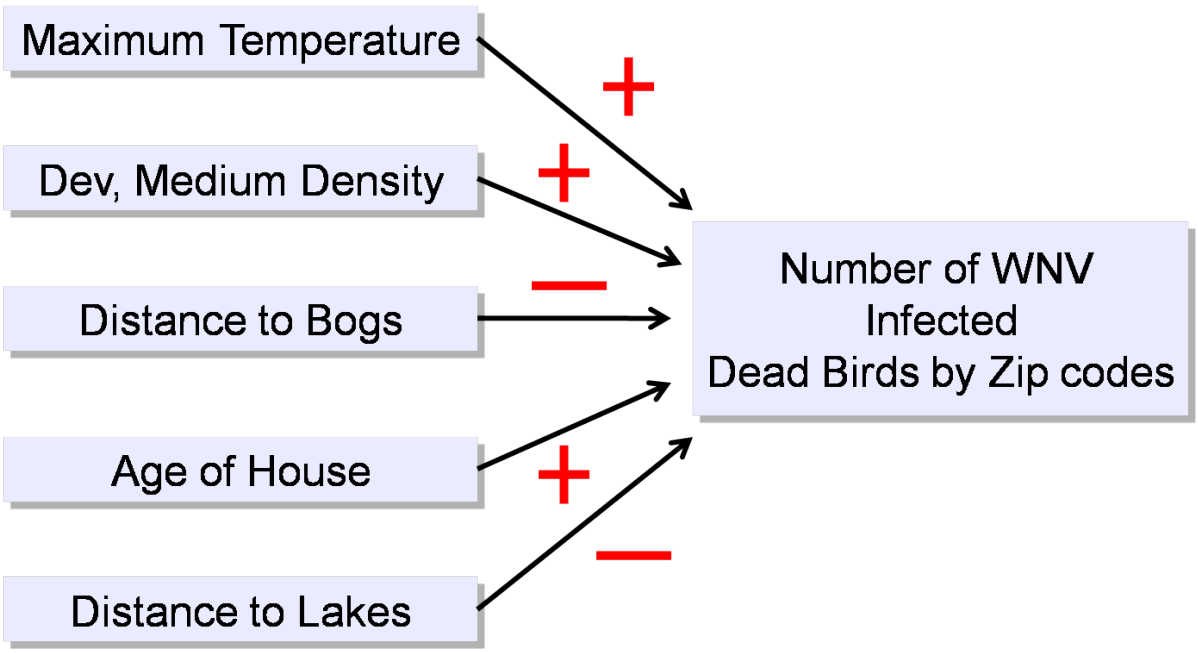


Figure 57 Visualizing the positive relationship between maximum daily temperature and West Nile virus infected dead birds by zip codes

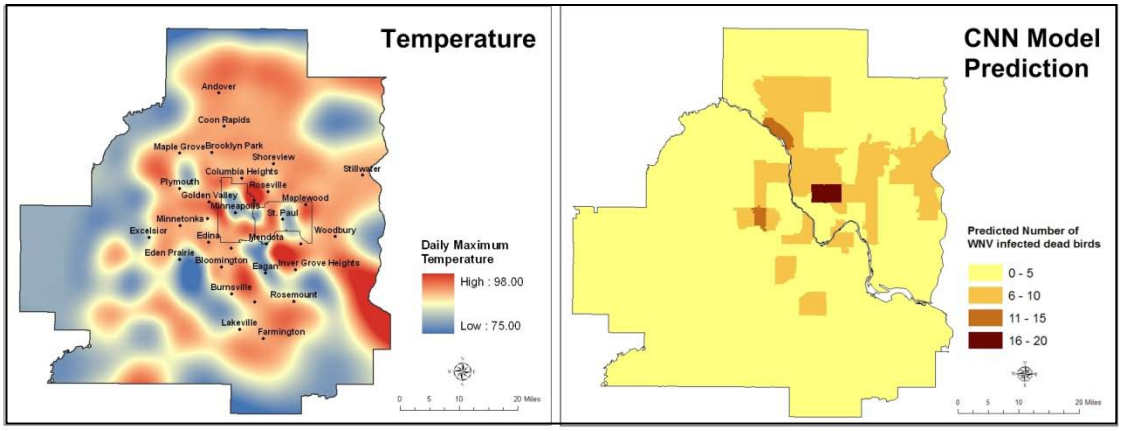


Figure 58 Visualizing the negative relationship between distance to bogs and West Nile virus infected dead birds by zip codes

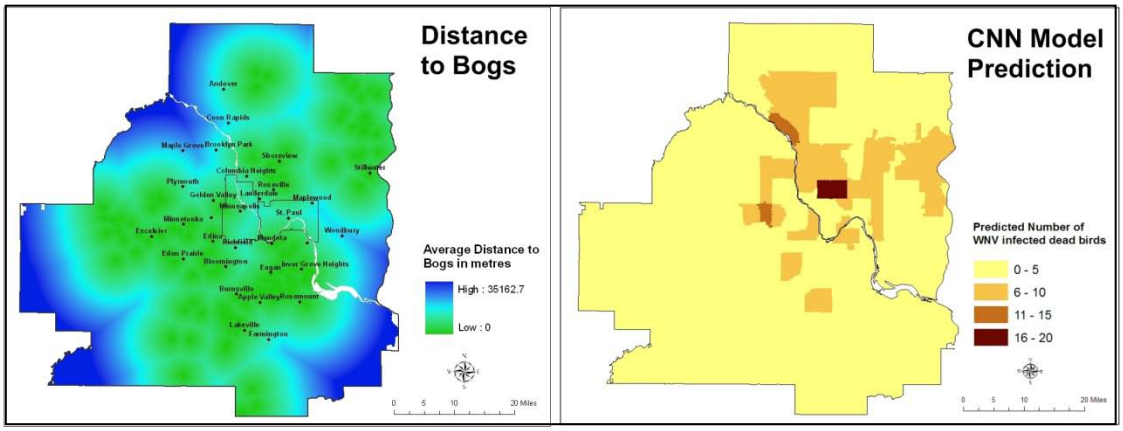


Figure 59 Visualizing the positive relationship between age of houses and West Nile virus infected dead birds by zip codes

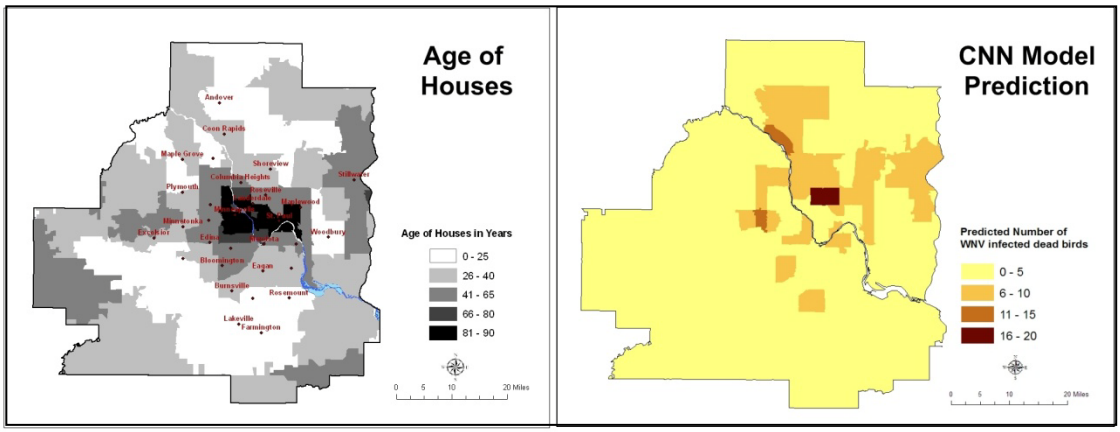


Figure 60 Visualizing the positive relationship between developed medium density land cover and West Nile virus infected dead birds by zip codes

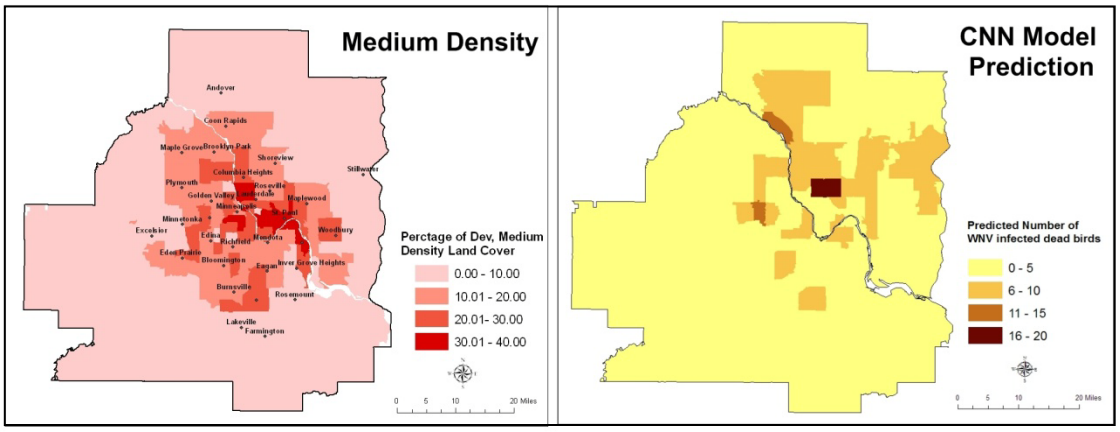
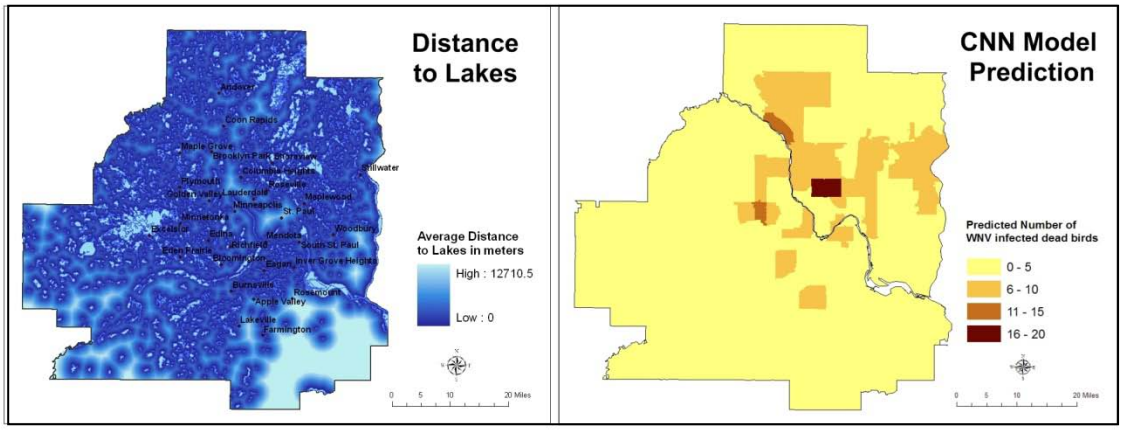


Figure 61 Visualizing the negative relationship between developed medium density land cover and West Nile virus infected dead birds by zip codes



In summary, the risk factors as well as their correlations that played major roles in the predicted value of infected dead birds encoded by the CNN model were similar to that of the linear model. However there were a few differences which needed further explanation. The main difference was the order of importance of the risk factors. For instance, the linear model indicated that the developed medium density land cover played a very important role in explaining the occurrence of WNV infection in birds, whereas the CNN model accorded it a less significant role. On the other hand, distance to bogs, a wetland type with high potential for mosquito production and bird habitats, played a significant role in the CNN model. Similarly, the age of houses had moved up in the order of importance over the land cover variable in the CNN model. These two predictors characterized urban features and are both positively related to the WNV infection. Another important difference between the two models was the correlation of distance to lakes. The linear model interpreted the correlation to be positive, whereas the CNN model, as hypothesized, defined the correlation between distance to lakes and the occurrence of WNV infection as negative. These differences are not surprising,

since the CNN model correlated the predictor variables or the risk factors in a nonlinear fashion. Thus it is expected that the relative roles played by each of the risk factor in a nonlinear fashion encoded by the CNN model will be realistic and more accurate when compared to the linear model. The point to note here is that the reliability of results obtained from a nonlinear model, capturing the dynamics of a complex health outcome such as WNV, is higher than the results obtained from a linear model.

7.5 Discussion

Both the broad and the detailed interpretation techniques identified daily temperature as the most important risk factor for WNV infection in birds. The predictor importance plot (Figure 55) obtained from the broad interpretation showed that the highest relative difference of RMSE was associated with temperature. That is, scrambling the values of this predictor variable obscured the correlation with the infected number of dead birds and therefore led to the greatest loss of predictive power (as measured by the RMSE value). This indicated that temperature was the most important risk factor. The resultant combined weight matrix from the detailed interpretation (Table 35) also showed that temperature was the highest weighted input variable under both the first and the second hidden layer neurons. The high values and positive signs of the weights indicated a strong positive correlation between maximum daily temperature and infection in birds. This finding was in tune with the previous studies (Ruiz et al. 2004; Hayes et al. 2005; Shaman et al. 2005; Gibbs et al. 2006; Vaidyanathan and Scott 2006; Adlouni et al. 2007; Gleiser et al. 2007; Landesman et al. 2007; Zou et al. 2007; Bouden et al. 2008).

This study made a significant contribution to the literature of WNV by investigating the relationship between WNV disease occurrences and *distance to* hypothesized risk factors, such as bogs and lakes. Other studies have quantified these

risk factors as percentages of area or presence and absence in an unit (Reisen et al. 2000; Schafer et al. 2004; Williot 2004; Cooke, Grala et al. 2006; Ezenwa et al. 2007; Leblond et al. 2007). On the contrary, this study attempted to analyze an important question, i.e., how does proximity or *spatial distribution* of a particular risk factor influences the risk of WNV. Bog is one of the major wetland types classified by MMCD as potential breeding site for mosquitoes. It is described as a wetland with vegetation including moss, sedges, cotton grass, and shrubs, poorly drained soil, waterlogged, and supporting a covering of plant residues (see Chapter 2 for details). Both the broad and the detailed interpretation techniques indicated distance to bog as the second most contributing factor with strong negative correlation with WNV disease incidences. Thus areas near bogs or with deceasing proximity to bogs the risk of WNV infection increased. The CNN model also reported a negative correlation between distance to lakes and infection in birds but the role played by it was not as important as the other variables.

The broad interpretation ranked housing age as the third most important factor. The detailed interpretation with positive combined weights indicated a positive relationship between age of houses and WNV infection. Only one other study, based in Chicago and its surrounding area, reported such relationship (Ruiz et al. 2004; Ruiz et al. 2007). The positive correlation between older houses and WNV activities could be explained by two plausible reasons. The first reason was due to maintenance of older houses. Typically, flooded and damp basements in older houses, cracks in the walls, open containers with stagnating water, shrubs in the backyards can provide suitable breeding conditions for mosquitoes. The second reason was the timing or the period during which these houses were built. The Chicago study showed positive association between houses built during the post World War II period from 1940 – 1960 and WNV activities (Ruiz et al. 2007). It was possible that with sudden increase in population and demand for housing, less attention was given to the location, surrounding drainage structure, physiographic, and soil characteristics of the areas in which houses were built. This lack of engineering foresight often resulted in houses with flooded basements. This

issue was then quickly mitigated by constructing catch basins very near to the houses, which can accumulate urban run-off water. However if these catch basins were left unattended, the combination of standing water and organic matter could likely provide potential breeding grounds for *Culex* mosquitoes and hence increase the risk of WNV occurrences. This explanation is also applicable in TCMA, where higher density of urban catch basins and storm water ponds are found in areas with higher numbers of houses 40 to 60 years old, probably built during the post World War II period. Further, the fact that Chicago and the TCMA have similar geologic features formed due to the glacial drift and massive retreat of ice sheets about 75,000 year ago, resulting in glacial deposits, erosion, lakes, and flat terrains with sandier soils added more confidence to this explanation.

The positive correlation between developed medium density land cover class and occurrence of WNV infection in birds supported the urban/suburban nature of disease transmission in the TCMA described in Chapter 4 and Chapter 5. Both the chapters with preliminary t-tests (Chapter 4) and Principal Component Analysis (Chapter 5) hypothesized that the dynamics of WNV illness in birds, mosquitoes, and humans in the TCMA show strong association with urban landscape features. Some of the urban characteristics associated with disease incidence were developed land cover classes (developed open space, developed high, medium, and low density), higher housing density, houses 40 to 60 years old, higher density of urban catch basins and storm water ponds, and smaller distances to pockets of natural areas within the city namely lakes, bogs, open green space, swamps, and trails. This urban preference of WNV transmission hypothesized by exploratory data analysis was confirmed rigorously by a nonlinear WNV model. These findings suggested that *Culex restuans* and *Culex pipiens*, which are predominantly urban mosquitoes, might play a major role in the transmission and amplification of the virus. The urban centric nature WNV occurrence in the TCMA reflected similar trends of virus transmission in the northeastern United States (Brown et al. 2008), Chicago during the 2002 WNV outbreak (Ruiz et al. 2004; Ruiz et al. 2007), Detroit metropolitan area (Ruiz et al. 2007), and Georgia (Gibbs et al.

2006).

7.6 Vector Control Policy Recommendations

Effective vector control programs to prevent and regulate the occurrence of West Nile virus incidences is dependent upon integration of two fundamental questions, *when* and *where* should we apply insecticides and pesticides to eliminate larva and adult mosquitoes. The research findings from this chapter can significantly answer the *where* question. The extracted relationships between the risk factors and WNV infection can identify factors which can be contained, regulated, or treated to reduce the risk of WNV. The results can also contribute to identifying risk areas for targeted preventive and control programs.

Typically comprehensive vector control and prevention strategies include surveillance of WNV activities, source reduction of mosquitoes including sanitation and water management guidelines, chemical controls by larviciding and adulticiding, health education, and public awareness on basic precautions against WNV (CDC 2003). The relationships that emerged between the risk factors (distance to bogs, lakes, developed medium density land cover, and housing age) and occurrence of WNV infection in birds from the CNN model could be incorporated in some of the mosquito abatement programs in the TCMA. These vector control measures are *source reduction* and *chemical control* programs.

CDC defines source reduction as a technique to eliminate habitats for mosquito larval. This is the most effective and economical method of providing long-term mosquito control. Source reduction can include activities such as disposal of used tires, cleaning of rain gutters, maintaining swimming pools by individual property owners, and extensive regional water management projects. These programs are typically conducted by mosquito control agencies, for example, Metropolitan Mosquito Control

District (MMCD) in the TCMA. All of these activities eliminate or substantially reduce mosquito breeding habitats as well as the need for repeated applications of insecticides in the affected habitats. Source reduction activities can be separated into two general categories: sanitation and water management. The WNV model indicated that there was a positive correlation between the infection and areas with developed medium density land cover, which is mostly occupied by single family residential houses. Therefore to prevent and contain further spread of WNV in the TCMA, specific and structured sanitation guidelines can be issued by the local or state health departments for individual property owners. Guidelines of basic precautionary measures such as tire removal, catch-basin cleaning, container removal, and maintenance of grassy backyards and swimming pools can be effective in reducing mosquito population. Some of the sanitation problems caused due to neglect, oversight, or lack of information on the part of property owners can be resolved through inspections by the local health or mosquito surveillance agency inspectors. Educational information about the importance of sanitation in the form of videos, slide shows**,** and fact sheets distributed at press briefings, fairs, schools and other public areas can also be effective.

Another finding from the model interpretation was that areas around older houses (40 to 60 years old) with higher density of catch basins and storm water ponds in close proximity were at higher risk for WNV activities. Implementation of source reduction strategies at catch basins and storm water structures can play an important role here. These man-made catch basins are constructed to hold urban runoff water before it is discharged into groundwater or surface water. However, over time and with lack of inspection, water stagnates in these catch basins and often with thick mat of decaying leaves. These conditions are favorable for mosquito production especially *Culex pipiens*, one of the main carrier of WNV in Minnesota. Therefore extensive mapping, regular inspection, treatment, and cleaning of catch basins and storm water ponds are crucial in eliminating such habitats. Regular and planned pest control treatment of older houses with flooded and damp basements can also be productive in reducing vector population.

This study also found a strong negative relationship between proximity to bogs and occurrence of WNV infection in birds, i.e., with closer proximity to bogs there is a higher risk of WNV infection. In addition, the WNV model also indicated relatively weaker negative relationship between distance to lakes and disease occurrence. These findings can be very useful for targeting chemical control vector abatement programs, such as larviciding and adulticiding. Larviciding is the application of chemicals to kill mosquito larvae or pupae by ground or aerial treatments, and is typically more effective in target-specific applications (CDC 2003). The objective of larviciding is to control the immature stages at the breeding habitat before adult populations have had a chance to disperse and therefore accuracy of application is important because missing even a relatively small area can cause the emergence of a large mosquito brood resulting in the need for broad-scale adulticiding. Adulticiding, on the hand, is the application of pesticides to kill adult mosquitoes (CDC 2003). This approach is a practical solution to control vector populations in situations of quick reduction of adult mosquito population to lower the risk of WNV transmission to humans. Hence adulticiding can be an effective strategy to reduce vector populations near bogs and lakes to lower the risk of WNV transmission to humans especially during increased recreational activities in and around water features.

7.7 Conclusion

This chapter has both methodological and applied contributions to the field of health geography and public health. In the following paragraphs, I will first discuss the methodological contributions in terms of interpreting neural network model to understand the dynamics of WNV transmission in the TCMA. Second, the applied contributions will recommend vector control policies to prevent the future spread of WNV in TCMA.

I have used two interpretation techniques, broad and detailed, to interpret the

relationships of the input risk factors and the occurrence of WNV infected dead birds encoded by the WNV neural network model. The broad interpretation technique was able to rank the relative importance of risk factors. The representation of risk factor importance plot allowed easy visualization of the predictors or risk factors in the model. In addition, apart from quantitative rankings, the plot provided a qualitative view of how important a given predictor was relative to others by looking at the difference between RMSE represented by the X-axis. Even though the broad interpretation approach provided an easy interpretation of predictor importance in explaining the response variable, it did not allow the user to elucidate exactly how a given risk factor affected the model output. That is, the methodology did not indicated the sign (or direction) of the effect of each input risk factor. Therefore I could not draw conclusions regarding the nature of correlations between the input risk factors and the occurrence of WNV incidences. For this I turned towards the detailed interpretation.

The detailed interpretation method based on weights and biases flowing down from the input layer to the output layer provided means for understanding the relationships between input risk factors and occurrences of WNV incidences. The methodology is similar to the PLS interpretation method for linear regression model. The analogy to the PLS method is strengthened when we consider that the hidden layer neurons are analogous to latent variables or components. The analysis of combined weights and the signs associated with each of the input factors in the weight matrix allowed deciphering the relationships between the predictor variables and the neural network output. The detailed interpretation of WNV CNN model indicated that maximum daily temperature, age of houses, and developed medium density land cover were positively related and distance to bogs and lakes were negatively related to the incidence of WNV infection in birds.

In summary, the risk factors and their correlations were similar to that of the OLS model, however, some of the differences indicated that the neural network model was better suited to capture the nonlinear relationships with greater accuracy than the linear regression model. The CNN interpretation methodologies presented in this

chapter provide a means for using CNN models for both predictions as well as for understanding the relationships present in the dataset. The methods are quite general as it requires only input data for the broad technique and optimized weights and biases from the network for the detailed approach. Hence they can be applied to other types of neural network algorithms, namely back propagation and feed forward. In summary, these methods expand the role of CNN models in social and health sciences as both predictive and explanatory tools, thus alleviating the black box nature of the neural network methodology to some extent.

In terms of applied contributions, the findings from this chapter could be used for vector control and prevention recommendations. In summary the suggested vector control policies are as follows:

1. Provide guidelines to individual property owners residing in the developed medium density area. Basic sanitation procedures will include maintenance of houses, their surroundings, and reducing probable mosquito habitats.
2. Educational information on the importance of sanitation in the form of videos, and fact sheets distributed at fairs, schools and other public areas can also be effective.
3. Extensive mapping, regular inspection, treatments, and cleaning of catch basins and storm water ponds can be crucial in eliminating favorable mosquito habitats
4. Regular and planned pest control treatment of older houses
5. Targeting chemical control vector abatement programs, such as larviciding and adulticiding in areas around bogs and lakes.

8. Chapter 8: Conclusion

8.1 Summary

The WNV is an infectious disease spreading rapidly throughout the United States, causing illness among thousands of birds, animals, and humans. Yet, we only have a rudimentary understanding of how the mosquito-borne virus operates in complex avian-human-environmental systems. The focus of my dissertation research is to develop novel approaches to explore, predict, and understand the disease by using key environmental, built environment, and anthropogenic risk factors that determine *why*, *when*, and *where* WNV strikes in the Twin Cities Metropolitan area (TCMA) of Minnesota. The techniques allowed me to answer four research questions, which are as follows: (1) Where are the exposure areas of West Nile virus in the TCMA? (2) Does the spatial distribution of WNV infected dead birds, positive mosquito pools, and human cases show clustering around the cities of Minneapolis and Saint Paul? (3) How do urban landscape features in the TCMA associate with the transmission of the virus? (4) What are the contributing risk factors and how are they related to the occurrence of WNV incidences in the TCMA? The preceding chapters investigated these research questions in detail including background review of literature, methodology used, results obtained, and analysis of findings for vector control policies. The research findings of this dissertation contributed to two fundamental questions of vector control strategies, *when* and *where* we should apply insecticides and pesticides to reduce the risk of WNV. The following paragraphs give a brief summary of the chapters.

Chapter 1 briefly traced the geographical spread of WNV in the United States

from its epicenter in New York in 1999 to the present scenario in 2008. Since its initial occurrence WNV has spread rapidly south and west causing seasonal epidemics and illness among thousands of birds, mosquitoes, humans, and animals. It is also the largest epidemic of human West Nile neuroinvasive disease (WNND) to date in North America (Marfin 2001; Huhn et al. 2002; Peterson et al. 2003; O'Leary et al. 2004). Between 1999 and 2001, few human cases were reported but by the end of 2008, the numbers increased dramatically to a total of 28,961 cases, of which 11,753 were classified as WNND, 16,463 as WNF, and 745 were unspecified clinical cases. By the end of 2008, the number of human fatalities due to WNV was 1,131 cases. The virus first reached Minnesota in 2002 and by 2003 WNV infections resulted in several regions of special epidemiological concern including the TCMA. The 2007 saw one of the severest incidences of WNV in Minnesota and by 2008, there were a total of 1598 infected dead birds and 473 human cases. Highest intensity of infections occurred in 2003, 2006, and 2007. The TCMA had consistently exhibited strong spatial clustering of WNV infected dead bird, mosquito, and human cases from 2002 to 2008, which reflected the similar pattern of WNV outbreaks in other urban areas of the country.

Chapter 2 focused on two preliminaries, the study area and the database. The WNV database stored, updated, and maintained georeferenced data on infected dead birds, positive mosquito pools, infected human cases, and hypothesized risk factors categorized into four broad groups of environmental, built-environment, proximity, and existing vector control measures.

Chapter 3 demonstrated the use of Nearest-Neighbor-Distance-Tome model (NNDT), as a new approach to delineate WNV exposure areas. The NNDT model uses GIS techniques and ecology of the virus to retrospectively delineate transmission cycles as exposure areas in their entirety, involving dead birds, mosquito pools, and human cases. The methodology is based on the combination of distance-time interaction and the temporal sequence of virus transmission from one component to another. This approach improved upon the use of arbitrary “critical” density measures, went beyond the use of clusters of a single component of the WNV transmission cycle, and provided

more flexibility in some respects than statistical models based on *a priori* relationships. More broadly, NNDT demonstrated how examining distances among the locations of infected dead birds, positive mosquito pools, and infected human cases infused a 'local' dimension that was combined with the 'global' knowledge of the time required for the virus to be transmitted from one component to another in the cycle. Temporal analysis of the 14 transmission cycles delineated by the model showed evidence of substantial reduction in the reporting of dead birds after the 10 -15 days window either due to localized reduction in bird population or migration or both. This indicated a signal of reduction in transmission activities due to lack of abundant reservoir (bird) population and virus proliferation ceased further with the onset of human illness. This finding can contribute to vector control strategies. The policy implications from this finding are explained in the section 8.2. Despite the success of the NNDT methodology, certain issues require further research and are worthy of mention. The NNDT technique is sensitive to the number of WNV incidences and performs more efficiently with higher number cases and as such is dependent upon the efforts of surveillance programs. Also, the technique will have better results if the human cases are aggregated at much finer spatial resolution than the zip codes or at exact location.

Chapter 4 conducted an exploratory analysis of the correlation between WNV disease occurrence and the hypothesized risk factors with *t-test*. The results indicated two major findings which formed the basis for the later chapters. First, the mean values of all the weather variables were statistically different in zip codes with reported human cases than in zip codes with no human cases. The positive *t-values* indicated a positive association between weather variables (temperature and precipitation) and disease occurrence, which was similar to findings from other studies. Second, the *t-test* results also showed that several urban landscape features, such as developed land cover classes (developed open space, developed high, medium, and low density), higher housing density, houses built 50 to 60 years before, higher density of urban catch basins, and shorter distances to pockets of natural areas within the city namely lakes, bogs, open green space, swamps, trails, etc were associated with WNV infection in humans. This

finding hypothesized that the transmission of WNV in the TCMA is urban-centric and that *Culex restuans* and *Culex pipiens*, which are predominantly urban mosquitoes, might play a major role in the transmission and amplification of the virus.

Building upon the analysis in Chapter 4, Chapter 5 explored how *only* urban landscape features contributed to the WNV activities in the TCMA. Following a three-step methodology of Principal Component and hierarchical cluster analysis with environment, built-environment, and proximity variables, the TCMA was divided into five urban landscape classes. The City-High density class, including the core urban areas of Minneapolis and Saint Paul, and the City-Medium density class, covering the immediate suburbs, reported high incidence rates of infected dead birds, mosquitoes and human cases. In addition specific urban features, such as, catch basins, housing density, age of houses, and proximity to lakes, swamps, bogs, and parks present within the urban areas were associated to the transmission of WNV virus. These results indicated that the Twin Cities urban landscape, a combination of natural and man-made features, provided suitable bird and mosquito habitats. These research findings were in tune with the urban centric nature of WNV transmission in Chicago and Detroit and therefore strengthened the general hypotheses that WNV infection had exhibited strong spatial clustering in the urban areas of Midwestern United States. This investigation also contributed to the broader research question in the field of health geography and public health, of how the heterogeneous urban landscape affects human health and disease pattern.

Chapter 6 presented in details the steps involved in developing a nonlinear computational neural network (CNN) model capturing the complex relationships between the risk factors and WNV disease occurrence in birds. The model output predicted the number of WNV infected dead birds by zip codes for the entire seven counties area of TCMA. Initially Genetic Algorithm (GA) was used to search the predictor space for the best subset of predictor variables which were then included in a CNN model. The final model was an optimized model with a lower RMSE value, higher R^2, and lowest difference between R^2 and Q^2. This procedure resulted in a 5-2-1 CNN architecture as the best model for WNV analysis with RMSE of 1.78, and R^2 and

Q^2 values of 0.75 and 0.62 respectively. The predictor variables included were distance to bogs, distance to lakes, daily maximum temperature, housing age, and percentage of developed medium density land cover class. This chapter also presented a comparative analysis between the CNN model and OLS model with same specifications. Even though the OLS model was statistically significant (*F*-statistic), the RMSE (2.56) was significantly higher and R^2 (0.51) was lower than the values obtained from the CNN model. The observed verses predicted histograms and the choropleth maps suggested that the CNN model had superior predictive capabilities than the OLS model when considering the dynamics of WNV. The external cross-validation results with new datasets also indicated that the CNN model had better predictive capabilities.

Chapter 7 had both methodological and applied contributions to the field of health geography and public health. Here I discuss the methodological contributions in terms of interpreting neural network models. The discussion on vector control policy implications are presented in section 8.2. I have adapted two interpretation techniques, broad and detailed, to interpret the relationships of the input risk factors and the occurrence of WNV infected dead birds encoded by the WNV neural network model. The broad interpretation technique was able to rank the relative importance of risk factors, which in descending order were average daily maximum temperature, distance to bogs, housing age, medium density land cover class, and distance to lakes. The representation of risk factor importance plot allowed easy visualization of the predictors or risk factors in the model. In addition, apart from quantitative rankings, the plot provides a qualitative view of how important a given predictor was relative to others by looking at the gap between the predictors on the X-axis. If there was a large gap or separation, for example between average daily maximum temperature and distance to bogs, one may determine that temperature plays a much more significant role than distance to bogs in the model's predictive power. We conjecture that predictors with little separation along the X-axis play similar roles within the CNN architecture, for example, distance to bogs, housing age, and medium density land cover. Even though the broad interpretation approach provided an easy interpretation of predictor

importance in explaining the response variable, it did not allow the user to elucidate exactly how a given risk factor affected the model output. That is, the methodology did not indicate the sign (or direction) of the effect of input risk factors. For this we turn towards the detailed interpretation.

The detailed interpretation method based on weights and biases, flowing down from the input layer to the output layer in a neural network, provided a means for understanding the relationships between input risk factors and occurrences of WNV incidences. The methodology is similar to the PLS interpretation method for linear regression model. The analogy to the PLS method is strengthened when we consider that the hidden layer neurons are analogous to latent variables or components in a PLS analysis. The analysis of combined weights and the signs associated with each of the input factors in the weight matrix allowed deciphering the relationships between the predictor variables and the neural network output. The detailed interpretation of WNV CNN model indicated that maximum daily temperature, age of houses, and developed medium density land cover class were positively related and distance to bogs and lakes were negatively related to the incidence of WNV infection in birds. In overall, the types of risk factors and their correlations were similar to that of the OLS model, however, some of the critical differences indicated that neural network model were better suited to capture nonlinear relationships with greater accuracy than the traditional linear regression model. The CNN interpretation methodologies presented in this chapter provided a means for using CNN models both for prediction as well as for understanding the relationships present in the dataset. The methods are quite general as it requires only input data for the broad technique and optimized weights and biases from the network for the detailed approach. Hence they can be applied to other types of neural network algorithms. In summary these methods expands the role of CNN models in social and health sciences as both predictive and explanatory tool, thus alleviating the black box nature of the neural network methodology to some extent.

8.2 Vector control policy implications

The mosquito control programs practiced by the Metropolitan Mosquito Control District (MMCD) of Minnesota targets the summer pest mosquito *Aedes vexans*, several species of spring *Aedes*, the cattail mosquito *Coquillettidia perturbans*, the eastern treehole mosquito *Aedes triseriatus* (La Crosse encephalitis vector), and the vector of western equine encephalitis *Culex tarsalis* (MMCD 2004). The arrival of WNV in Minnesota in 2002 elevated the importance of controlling *Culex pipiens* and the three other *Culex* species (*Culex tarsalis, Cules restauns,* and *Culex salinarius*). Larval control is the main focus of the program but is supplemented by adult mosquito control when necessary. However these mosquito abatement measures can be further improved by integrating the answers of two fundamental questions to eliminate larva and adult mosquitoes. First, *when (time)* should we apply insecticides and pesticides? Second, *where (area)* should we target spraying of pesticides or source reduction programs. The research findings from this thesis can significantly contribute in answering both the questions.

In Chapter 3 NNDT results indicated that dead bird reports were an essential part of the delineation of WNV cycles at localized scale. In the TCMA, the transmission cycle peaked 10-15 days prior to the onset of human illness, a period that was consistent with the epidemiology of the virus transmission from bird to mosquito and then to humans. Temporal analysis of the 14 delineated cycles at local scales showed evidence of substantial reduction in the reporting of dead birds after the 10 -15 days window either due to localized reduction in bird population or migration or both. This indicated a signal of reduction in transmission activities due to lack of abundant reservoir (bird) population and virus proliferation ceased further with the onset of human illness. These findings can contribute to the *when* question mentioned above. For example, adulticiding may be best concentrated during the temporal window of 10-15 days after the reporting dates of few WNV-infected dead birds because by the time humans develop WNV symptoms in a given area, the risk of further amplification of WNV disease may have subsided and thus widespread spraying of pesticides will have little or

no effect. This would also allow efficient allocation of resources and balance between the need for mosquito eradication and desire to limit the environmental impacts from unwarranted pesticide usage.

These findings are also expected to influence WNV-related legal responses that could further environmental health. As was done in New York City, legislators in the TCMA could waive the procedural requirements for obtaining a permit to apply insecticides and pesticides if the threat of a WNV outbreak is imminent in the delineated high-risk areas (Hodge and O'Connel 2005). This would result in prompt mosquito control and reduce the risk of WNV. In addition, this would also protect the environment from unwarranted use of pesticides in non-risk areas.

In Chapter 5, the TCMA was divided into five urban classes based on the combination of hypothesized WNV natural and built-environment risk factors. The derived urban classes were City-High Density in the center, surrounded by City-Medium Density, Suburb, Outer Suburb 1, and lastly Outer Suburb 2. The City-High density and City-Medium density classes reported high incidence rates of infected dead birds, mosquitoes and human cases. This urban preference of WNV transmission in the TCMA indicates that MMCD should elevate the surveillance and treatment of potential habitats for *Culex restuans* and *Culex pipiens*. These species are predominantly urban mosquitoes and carriers of WNV in Minnesota. In addition, the derived urban landscape classes can provide a basis for the selection of field sites for mosquito traps and collection, testing, and treatment. This could be very useful for field staff at the MMCD.

The broad and the detailed interpretation of WNV neural network model indicated that age of houses and developed medium density land cover were positively related to the WNV infection and distance to bogs and lakes were negatively related to the incidence of WNV infection in birds (chapter 7). These extracted relationships between the risk factors and disease occurrence can contribute in identifying risk areas for targeted preventive and control programs and therefore answer the *where* question mentioned above. In summary the suggested vector control from these results policies are as follows:

1. Provide guidelines to individual property owners residing in the developed medium density area to follow basic sanitation procedures in maintaining the surrounding of houses and eliminating potential mosquito breeding sources.
2. Educational information about the importance of eliminating mosquito larva in the form of videos, and fact sheets distributed at fairs, schools and other public areas can also be effective.
3. Extensive mapping, regular inspection, treatments, and cleaning of catch basins and storm water ponds can be useful in eliminating favorable mosquito habitats.
4. Regular and planned pest control treatments of older houses
5. Targeting chemical control vector abatement programs, such as larviciding and adulticiding in areas around bogs and lakes.

8.3 Do home foreclosures worsen West Nile virus?

The recent mortgage crisis has severely affected the world's financial system but can also be a leading contributor to WNV. There are speculations from different parts of the country, such as California, New Orleans, Phoenix, Washington DC, etc that with significant rise in the number of foreclosed houses, incidences of WNV infections in birds, mosquitoes, and humans increased significantly. These recent events and speculations of local and state health workers led to another research topic associated to WNV. That is, whether there is a relation between rising foreclosures and increase in WNV activities or it is just a coincidence.

A dramatic increase in foreclosures, abandoned houses, and slow rate of home sales have dotted the urban and suburban landscape with higher number of unmowed lawns and stagnant water in swimming pools which are often covered with thick mat of decaying leaves. These new attractive breeding grounds for mosquitoes, especially urban species like *Culex pipiens* and *Culex restuans* could potentially increase the threat

of WNV epidemics. “*It's an example of how events seemingly unrelated to disease can impact public health*”, says Roger Nasci, PhD, chief of the CDC's arboviral disease department. This also emphasizes the fact that there are number of complicating and interrelated factors that at first might not seem a risk factor but can play an important role in the spread of the virus. For example, this foreclosure issue illustrates how changes in economic conditions affect the spread of WNV in birds, mosquitoes, and humans.

References

Adlouni, S. E., C. Beaulieu, T. B. Ouarda, P. L. Gosselin, and A. Saint-Hilaire. 2007. Effects of climate on West Nile Virus transmission risk used for public health decision-making in Quebec. *International Journal of Geographic Information Science* 6 (40).

Alberti, M. 2005. The effects of urban patterns on ecosystem function. *International Regional Science Review* 28:168 - 192.

Alberti, M., J. Marzluff, E. Shulenberger, G. Bradley, C. Ryan, and C. ZumBrunnen. 2003. Integrating humans into ecology: opputunities and challenges for studying urban ecosystems. *BioSience* 53:1169 - 1179.

Altman, D. G., and P. K. Anderson. 1989. Bootstrap investigation of the stability of a Cox regression model. *Statistcis in Medicine* 8:771 - 783.

Anderberg, M. R. 1973. *Cluster Analysis for Applications*. New York: Academic Press.

Andreadis, T., J. Anderson, and C. Vossbrinck. 2001. Mosquito surveillance for West Nile virus in Connecticut, 2000: isolation from Culex pipiens, Cx. restuans, Cx. salinarius, and Culiseta melanura. *Emerging Infectious Diseases* 7 (4):670 - 674.

Aramaki, T., M. Galal, and K. Hanaki. 2006. Estimation of reduced and increasing health risks by installation of urban wastewater systems. *Water Science and Technology* 53 (9):247--252.

Barrera, R., N. Torres, J. Freier, J. Navarro, C. Garcia, R. Salas, C. Vasquez, and S. Weaver. 2001. Characterization of enzootic foci of Venezuelean equine encephalitis virus in Western Venezuela. *Vector Borne Zoonotic Diseases* \ 1:219 - 230.

Bell, A., C. Brewer, N. Mickelson, G. Garman, and J. Vaughan. 2006. West Nile virus epizootiology, Central Red River Valley, North Dakota and Minnesota, 2002 - 2005. *Emerging Infectious Diseases* 12 (8):1245 - 1247.

Berman, M., and P. J. Diggle. 1989. Estimating Weighted Integrals of the Second-order Intensity of a Spatial Point Process. *Journal of Royal Statistical Society* B (51):81 - 92.

Berry, B. 1971. *City Classification Handbook: methods and applications* New York: Wiley.

Blunden, J. R., W. T. R. Pryce, and P. Dreyer. 1998. The classification of rural areas in the European context: An exploration of a typology using neural network applications. *Regional Studies* 32 (2):149-160.

Bolling, B. G., C. G. Moore, S. L. Anderson, C. D. Blair, and B. J. Beaty. 2007. Entomological studies along the Colorado front range during a period of intense West Nile virus activity. *Journal of the American Mosquito Control Association* 23 (1):37-46.

Bostwick, D. G., J. Adolfsson, H. B. Burke, J. E. Damber, H. Huland, M. Pavone-Macaluso, and D. J. Waters. 2005. Epidemiology and statistical methods in prediction of patient outcome. *Scandinavian Journal of Urology and Nephrology* 39:94-110.

Bouden, M., B. Moulin, and P. L. Gosselin. 2008. The geosimulation of West Nile virus propagation: a multi-agent and climate sensitive tool for risk management in public health. *International Journal of Health Geographics* 7 (35).

Bowman, C., A. B. Gumel, P. v. d. Driessche, J. Wu, and H. Zhu. 2005. A mathematical model for assessing control strategies against West Nile Virus. *Bulletin of Mathemetical Biology.* 67:1107-1133.

Brown, H. E., J. E. Childs, M. A. Diuk-Wasser, and D. Fish. 2008. Ecological factors associated with West Nile virus transmission, northeastern United States. *Emerging Infectious Diseases* 14 (10):1539-1545.

Browne, M., B. Castelle, D. Strauss, R. Tomlinson, M. Blumenstein, and C. Lane. 2007. Near-shore swell estimation from a global wind-wave model: Spectral process, linear, and artificial neural network models. *Coastal Engineering* 54 (5):445-460.

Brownstein, J. S., H. Rosen, D. Purdy, J. R. Miller, M. Merlino, and F. Mostashari. 2002. Spatial analysis of West Nile virus: rapid risk assessment of an introduced vector-borne zoonosis. *Vector Borne Zoonotic Disease.* 2:157-164.

Burke, H. B., P. H. Goodman, D. B. Rosen, D. E. Henson, J. N. Weinstein, F. E. Harrell, J. R. Marks, D. P. Winchester, and D. G. Bostwick. 1997. Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer* 79

(4):857-862.

CDC. 1999. Outbreak of West Nile like viral encephalitis - New York. In *MMWR Morb Mortal Wkly Rep*, 845-849: Centers for Disease Control and Prevention.

———. 2003. Epidemic/Epizootic West Nile Virus in the United States: Guidelines for Surveillance, Prevention, and Contro. Fort Collin, Colorado: Center for Disease Control and Prevention.

Chang, I. C., and T. Y. Hsiao. 2004. Short-term model of the production of construction aggregates in Taiwan based on artificial neural networks. *Environmental Science and Pollution Research* 11 (2):84-90.

Clair, T. A., and J. M. Ehrman. 1996. Variations in discharge and dissolved organic carbon and nitrogen export from terrestrial basins with changes in climate: A neural network approach. *Limnology and Oceanography* 41 (5):921-927.

Clark, C., L. Ryan, I. Kawachi, M. J. Canner, L. Berkman, and R. J. Wright. 2008. Witnessing community violence in residential neighborhoods: A mental health hazard for urban women. *Journal of Urban Health-bulletin of the New York Academy of Medicine* 85 (1):22--38.

Cook, R. D., and S. Weisberg. 1999. *Applied Regression Including Computing and Graphics*: Wiley.

Cooke, W. H., G. Katarzyna, and R. C. Wallis. 2006. Avian GIS models signal human risk for West Nile Virus in Mississippi. *International Journal of Health Geographies* 5 (36).

Cooke, W. I., K. Grala, and R. Wallis. 2006. Avian GIS models signal human risk for West Nile virus in Mississippi. *International Journal of Geographic Information Science* 5 (36).

Corrigan, R. L. A., C. Waldner, T. Epp, J. Wright, S. M. Whitehead, H. Bangura, E. Young, and H. G. G. Townsend. 2006. Prediction of human cases of West Nile virus by equine cases, Saskatchewan, Canada, 2003. *Preventive Veterinary Medicine* 76 (3-4):263-272.

Cruz, L., V. M. Cardenas, and A. M. 2005. Serological evidence of West Nile virus activity in EL Salvador. *American Journal of Tropical Medicine and Hygiene* 72:612 - 615.

David, S., S. Mak, L. MacDougall, and M. Fyfe. 2007. A bird's eye view: using geographic analysis to evaluate the representativeness of corvid indicators for West Nile virus surveillance. *International Journal of Health Geographics* 6 (1):3.

DeGroote, J. P., R. Sugumaran, S. M. Brend, B. J. Tucker, and L. C. Bartholomay. 2008. Landscape, demographic, entomological, and climatic associations with human disease incidence of West Nile virus in the state of Iowa, USA. *International Journal of Health Geographics* 7 (19).

Diggle, P. J. 1985. A kernel method for smoothing point process data. *Journal of Royal Statistical Society* C (34):138 - 147.

Diuk-Wasser, M. A., H. E. Brown, T. G. Andreadis, and D. Fish. 2006. Modeling the spatial distribution of mosquito vectors for West Nile virus in Connecticut, USA. *Vector-Borne and Zoonotic Diseases* 6 (3):283-295.

Drechsel, P., B. Keraita, P. Amoah, R. C. Abaidoo, L. Raschid-Sally, and A. Bahri. 2008. Reducing health risks from wastewater use in urban and peri-urban sub-Saharan Africa: applying the 2006 WHO guidelines. *Water Science and Technology* 57 (9):1461--1466.

Eidson, M., L. Karmer, and W. Snow. 2001. Dead bird surveillance as an early warning system for West Nile virus. *Emerging Infectious Diseases* 7:631 - 635.

Eidson, M., M. Miller, and L. Kramer. 2001. Dead crow densities and human cases of West Nile virus, New York State. *Emerging Infectious Diseases* 7 (662 - 664).

Eidson, M., K. Schmit, Y. Hagiwara, M. Anand, P. B. Backenson, I. Gotham, and L. Kramer. 2005. Dead Crow Density and West Nile Virus Monitoring, New York. *Emerging Infectious Diseases* 11 (9):1370-1375.

ESRI. http://www.esri.com/. Redlands, CA.

Ewing, R., R. Schieber, and C. Zeeger. 2003. Urban sprawl as a risk factor in motor vehicle occupant and pdestrain fatalities. *American Journal of Public Health* 93:1541 - 1545.

Ewing, R., T. Schmid, R. Killingsworth, A. Zlot, and S. Raudenbush. 2003. Relationship between urban sprawl and physical activity, obesity, and morbidity. *American Journal of Public Health* 18:47 - 57.

Ezenwa, V., L. Milheim, M. Coffey, M. Godsey, R. King, and S. Guptill. 2007. Land cover variation and West Nile virus prevalence: patterns, processess, and implications for disease control. *Vector Borne Zoonitc Diseases* 7 (2):173 - 180.

Faris, R., and W. Dunham. 1939. *Mental Disorders in Urban Areas: An ecological study of schizohrania and other pshycoses* Chicago, IL: University of Chicago Press.

Florida Statutes. 403.7095.

Forrest, S. 1993. Genetic Algorithms: Principles of Natural Selecyion Applied to Computation *Science* 261:872 - 878.

Fotheringham, S., and S. Wong. 1991. bThe modifiale areal unit problem in multivariate statistical analysis. *Environment and Planning A* 23:1025 - 1044.

Gary, T. L., S. A. Stark, and T. A. LaVeist. 2007. Neighborhood characteristics and mental health among African Americans and whites living in a racially integrated urban community. *Health & Place* 13 (2):569--575.

GeoDa. http://geodacenter.asu.edu/software/downloads.

Ghosh, D., and S. M. Manson. 2008. Robust Principle Component Analysis and Geographically Weighted Regression: Urbanization in the Twin Cities Metropolitan Area of Minnesota. *The URISA Journal* 20 (1):15 - 24.

Gibbs, S. E. J., M. C. Wimberly, M. Madden, J. Masour, M. J. Yabsley, and D. E. Stallknecht. 2006. Factors affecting the geographic distribution of West Nile virus in Georgia, USA: 2002 - 2004. *Vector-Borne and Zoonotic Diseases* 6 (1):73 - 82.

Gleiser, R. M., A. J. Mackay, A. Roy, M. M. Yates, R. H. Vaeth, G. M. Faget, A. E. Folsom, W. F. Augustine, R. A. Wells, and M. J. Perich. 2007. West Nile virus surveillance in East Baton Rouge Parish, Louisiana. *Journal of the American Mosquito Control Association* 23 (1):29-36.

Golbraikh, A., and A. Tropsha. 2002. Beware of Q2. *Journal of Molecular Graphics and Modeling* 20:269 - 276.

Gordon, A. D. 1999. *Classification*. Second Edition ed. London: Chapman and Hall.

Greenberg, M., H. Mayer, K. Miller, R. Hordon, and D. Knee. 2003. Reestablishing public health and land use planning to protect public water supplies. *American Journal of Public Health* 93 (18):1522 - 1526.

Grineski, S. 2008. Coping with asthma in the central city: Parental experiences with children's health care. *Journal of Health Care For the Poor and Underserved* 19 (1):227--236.

Guha, R. 2005a. Methods to improve the Reliability, Validity, and Interpretability of QSAR Models Chemistry, Pennsylvania State University, State College, PA.

Guha, R., and P. Jurs. 2005b. Interpreting Computational Neural Network QSAR Models: A Measure of Descriptor Importance. *Journal of Chemical Information and Modeling* 45 (3):800-806.

Guha, R., D. Stanton, P. Jurs. 2005c. Interpreting Computational Neural Networks QSAR Models: A Detailed Interpretation of the Weights and Biases. *Journal of Chemical Information and Modeling* 45 (4):1109-1121.

Hall, T. 2006. *Urban Geography*. 3rd ed. London and New York: Routledge.

Hammad, T. A., M. F. AbdelWahab, N. DeClaris, A. ElSahly, N. ElKady, and G. T. Strickland. 1996. Comparative evaluation of the use of artificial neural networks for modeling the epidemiology of schistosomiasis mansoni. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 90 (4):372-376.

Hawthorne, B. Hawth's Analysis Tolls for ArcGIS. http://www.spatialecology.com/htools/tooldesc.php.

Hayes, C. G. 2007. West Nile Virus: Uganda, 1937, to New York City, 1999. *Annals New York Academy of Sciences*:25 - 37.

Hayes, E. B., N. Komar, R. S. Nasci, S. P. Montgomery, D. R. O'Leary, and G. L. Campbell. 2005. Epidemiology and Transmission Dynamics of West Nile Virus Disease. *Emerging Infectious Disease* 11 (8):1176-1173.

Haykin, S. 2001. *Neural Networks*. Singapore: Pearson Education.

Herold, M., H. Couclelis, and K. C. Clarke. 2005. The role of spatial metrics in the analysis and modeling of urban land use change. *Computers Environment and*

Urban Systems 29:369 - 399.

Hinckley, A. F., D. R. O'Leary, and E. B. Hayes. 2007. Transmission of West Nile virus through human breast milk seems to be rare. *Pediatrics* 119 (3):E666-E671.

Hirsch, S., J. Shapiro, and P. Frank. 1997. Use of an artificial neural network in estimating prevalence and assessing under diagnosis of asthma. *Neural Computing & Applications* 5 (2):124-128.

Hodge, J. G., and J. P. O'Connel. 2005. West Nile Virus: Legal Responses That Further Environmental Health. *Journal of Environmental Health.* 68 (1):44-47.

Howell, E., and J. McFeeters. 2008. Children's mental health care: Differences by race/ethnicity in urban/rural areas. *Journal of Health Care For the Poor and Underserved* 19 (1):237--247.

Hubalek, Z., and J. Halouzka. 1999. West Nile fever - a reemerging mosquito-borne viral disease in Europe. *Emerging Infectious Disease* 5:643 - 650.

Huhn, G., C. Austin, C. Langkop, K. Kelly, R. Lutch, R. Lampman, R. Novak, L. Haramis, R. Boker, and S. Smith. 2005. The emergence of west nile virus during a large outbreak in Illinois in 2002. *American Journal of Tropical Medicine and Hygeine* 72:768 - 776.

Huhn, G. D., C. Austin, C. Langkop, K. Kelly, R. Lycht, and R. Lampman. 2002. The emergence of West Nile virus during a large outbreak in Illinois in 2002. *American Journal of Tropical Medicine and Hygiene* 72:768 - 776.

Jaimes, F., J. Farbiarz, D. Alvarez, and C. Martinez. 2005. Comparison between logistic regression and neural networks to predict death in patients with suspected sepsis in the emergency room. *Critical Care* 9 (2):R150-R156.

Johnson, G. D. 2008. Prospective spatial prediction of infectious disease: experience of New York State (USA) with West Nile Virus and proposed directions for improved surveillance. *Environmental and Ecological Statistics* 15 (3):293-311.

Johnson, G. D., M. Eidson, K. Schmit, A. Ellis, and M. Kulldorff. 2006. Geographic prediction of human onset of West Nile virus using dead crow clusters: An evaluation of year 2002 data in New York State. *American Journal of Epidemiology* 163 (2):171-180.

Jones, A. Y. M., P. K. W. Lam, and M. D. I. Gohel. 2008. Respiratory health of road-side vendors in a large industrialized city. *Environmental Science and Pollution Research* 15 (2):150--154.

Kaplan , D., J. Wheeler, and S. Holloway. 2004. *Urban Geography*. New York: Wiley.

Kiang, R., F. Adimi, V. Solka, J. Nigro, P. Singhasivanon, J. Sirichaisinthop, S. Leemingsawat, C. Apiwathnasorn, and S. Looareesuwan. 2006. Meteorological, environmental remote sensing and neural network analysis of the epidemiology of malaria transmission in Thailand. *Geospatial Health* 1 (1):71-84.

Kitron, U. 1998. Landscape ecology and epidemiology of vector-borne diseases: tools for spatial analysis. *Journal of Medical Entomology* 35:435 - 445.

Koenig, W. D., L. Marcus, T. W. Scott, and J. L. Dickinson. 2007. West Nile virus and California breeding bird declines. *Ecohealth* 4 (1):18-24.

Komar, N., S. Langevin, S. Hinten, N. Nemeth, E. Edwards, D. Hettler, B. Davis, R. Bowen, and M. Bunning. 2003. Experimental infection of North American birds with the New York 1999 strain of West Nile virus. *Emerging Infectious Diseases* 9:311 - 322.

Kulldroff, M., and U. Hjalmars. 1999. The Knox method and other test for space-time interaction. *Biometrics* 55:544 - 552.

Kyrkilis, G., A. Chaloulakou, and P. A. Kassomenos. 2007. Development of an aggregate Air Quality Index for an urban Mediterranean agglomeration: Relation to potential health effects. *Environment International* 33 (5):670--676.

LaBeaud, A. D., A. M. Gorman, J. Koonce, C. Kippes, J. McLeod, J. Lynch, T. Gallagher, C. H. King, and A. M. Mandalakas. 2008. Rapid GIS-based profiling of West Nile virus transmission: defining environmental factors associated with an urban-suburban outbreak in Northeast Ohio, USA. *Geospatial Health* 2 (2):215-225.

LaDeau, S. L., A. M. Kilpatrick, and P. P. Marra. 2007. West Nile virus emergence and large-scale declines of North American bird populations. *Nature* 447 (7145):710-U13.

Landesman, W., B. Allan, R. Langerhans, T. Knight, and J. Chase. 2007. Inter-annual associations between precipitation and human incidence of West Nile virus in

the United States. *Vector Borne Zoonitc Diseases* 7 (3):337 - 343.

Lappo, G. M., N. V. Petrov, and A. John. 1992. *Urban Geography in the Soviet Union and the United States*: Rowman and Littlefield Publishers Inc.

Leblond, A., A. Sandoz, G. Lefebvre, H. Zeller, and D. J. Bicout. 2007. Remote sensing based identification of environmental risk factors associated with West Nile disease in horses in Camargue, France. *Preventive Veterinary Medicine* 79 (1):20-31.

Leventhal, T., and J. Brooks-Gunn. 2003. Moving to opportunity: an experimental study of neighborhood effects on mental health. *American Journal of Public Health* 83:1576 - 1582.

Lian, M., R. Warner, J. Alexander, and K. Dixon. 2007. Using geographic information systems and spatial and space-time scan statistics for a population-based risk analysis of the 2002 equine West Nile epidemic in six contiguous regions of Texas. *International Journal of Health Geographics* 6 (1):42.

Livingstone, D., and D. Manallack. 1993. Statistics Using Neural Networks: Chance Effects. *Journal of Medicinal Chemistry* 36:1295 - 1297.

Lloyd, R., and R. L. Bunch. 2003. Technology and map-learning: Users, methods, and symbols. *Annals of the Association of American Geographers* 93 (4):828-850.

Malkinson, M., and C. Banet. 2002. The role of birds in the etiology of West Nile virus in Europe and Africa. *Curr Top Microbial Immunology* 267:309 - 322.

Marfin, A. A., Peterson, L.R., Edison, M., Miller, J., Hadler, J., Farello, C., et al. 2001. Widespread West Nile virus activity, eastern United States, 2000. *Emerging Infectious Disease* 7:730-735.

Matter, S., E. Edwards, and J. Laguado. 2005. West Nile virus antibodies in Columbian horses. *Emerging Infectious Disease* 11:1497 - 1498.

McIntosh, B. M., G. M. McGillivray, and D. B. Dickenson. 1968. Ecological studies on Sindbis and West Nile viruses in South Africa: Infection in a wild avian population. *South African Journal of Medical Science* 33 (105: 112).

MDH. *West Nile virus maps and statistics* Minnesota Department of Health 2003 [cited.

Available from http://www.health.state.mn.us/divs/idepc/diseases/westnile/.

———. *West Nile virus maps and statistics*. Minnesota Department of Health 2005 [cited. Available from http://www.health.state.mn.us/divs/idepc/diseases/westnile/.

MDH, M. D. o. H. *http://www.health.state.mn.us/* [cited.

Miramontes, R. J., W. Lafferty, B. Lind, and M. Oberle. 2006. Is agricultural activity linked to the incidence of human West Nile virus? *American Journal of Preventive Medicine* 30 (2):160 - 163.

MMCD. 2002. Are these bugging you? St.Paul: Metropolitan Mosquito Control district.

———. 2004. 2004 Operational Review & Plans for 2005. Annual Report to the Technical Adviosry Board. St. Paul, Minnesota: Metropolitan Mosquito Control District.

MMCD, M. M. C. D. http://www.health.state.mn.us/.

Mongoh, M. N., M. L. Khaitsa, and N. W. Dyer. 2007. Environmental and ecological determinants of West Nile virus occurrence in horses in North Dakota, 2002. *Epidemiology and Infection* 135 (1):57-66.

Mostashari, F., K. Martin, J. J. Hartman, J. R. Miller, and V. Kulasekera. 2003a. Dead bird clusters as an early system for West Nile virus activity. *Emerging Infectious Diseases* 9 (6):641 - 646.

———. 2003b. Dead bird clusters as an early warning system for West Nile virus activity. *Emerging Infectious Diseases* 9 (6):641 - 646.

Murray, K., S. Baraniuk, M. Resnick, R. Arafat, C. Kilborn, K. Cain, R. Shallenberger, T. L. York, D. Martinez, J. S. Hellums, D. Hellums, M. Malkoff, N. Elgawley, W. McNeely, S. A. Khuwaja, and R. B. Tesh. 2006. Risk factors for encephalitis and death from West Nile virus infection. *Epidemiology and Infection* 134 (6):1325-1332.

NCDC. *National Climatic Data Center* [cited. Available from http://www.ncdc.noaa.gov/oa/mpp/freedata.html#TOP.

O'Leary, D., A. Marfin, S. Montgomery, K. AM., J. Lehman, and B. Biggerstaff. 2004. The epidemic of West Nile virus in the United States, 2002. *Vector Borne Zoonotic Disease.* 4:61-70.

OIE. 2004. West Nile fever in Belize: World Organization for Aninal Health.

Orme-Zavaleta, J., J. Jorgensen, B. D'Ambrosio, E. Altendorf, and P. A. Rossignol. 2006. Discovering spatio-temporal models of the spread of West Nile virus. *Risk Analysis* 26 (2):413-422.

Oudinet, J. P., J. Meline, W. Chehmicki, M. Sanak, D. W. Magdalena, J. P. Besancenot, S. Wicherek, B. Julien-Laferriere, J. P. Gilg, H. Geroyannis, A. Szczeklik, and K. Krzemien. 2006. Towards a multidisciplinary and integrated strategy in the assessment of adverse health effects related to air pollution: The case study of Cracow (Poland) and asthma. *Environmental Pollution* 143 (2):278--284.

Owen, S., A. MacKenzie, R. Bunce, H. Stewart, R. Donovan, G. Stark, and C. Hewitt. 2006. Urban land classification and its uncertainties using principle component and cluster analyses: A case study for the UK West Midlands. *Landscape and Urban Planning* 78:311 - 321.

Ozdenerol, E., E. Bialkowska-Jelinska, and G. N. Taff. 2008. Locating suitable habitats for West Nile Virus-infected mosquitoes through association of environmental characteristics with infected mosquito locations: a case study in Shelby County, Tennessee. *International Journal of Health Geographics* 7 (12).

Pan, L. L., L. X. Qin, S. X. Yang, and J. P. Shuai. 2008. A neural network-based method for risk factor analysis of West Nile virus. *Risk Analysis* 28 (2):487-496.

Peterson, L. R., A. A. Marfin, and D. J. Gubler. 2003. West Nile virus. *JAMA* 290:524 - 528.

Pradier, S., A. Leblond, and B. Durand. 2008. Land cover, landscape structure, and West Nile virus circulation in southern France. *Vector-Borne and Zoonotic Diseases* 8 (2):253-263.

PRISM. PRISM Group of Oregon State University. http://www.prism.oregonstate.edu/.

Quirin, R., R. Salas, and S. Zientara. 2004. West Nile virus. *Emerging Infectious Disease* 10:706 - 708.

R. The Comprehensive R Archive Network. http://cran.r-project.org/.

Rappole, J. H., S. R. Derrickson, and Z. Hubalek. 2000. Migratory Birds and spread of West Nile Virus in the Western Hemisphere. *Emerging Infectious Disease* 6 (4):319-328.

Reisen, W., H. Lothrop, R. Chiles, M. Madon, C. Cossen, L. Woods, S. Husted, V. Kramer, and J. Edman. 2004. West Nile virus in California. *Emerging Infectious Diseases* 10 (8):1369 - 1378.

Reisen, W. K., J. O. Lundstorm, T. W. Scott, and B. F. Eldridge. 2000. Patterns of avian seroprevalence to Western Equine Encephalomyelitis and Saint Louis Encephalitis viruses in California, USA. *Journal of Medical Entomology* 37 (507 - 527).

Rey, J. R., N. Nishimura, B. Wagner, M. A. H. Braks, S. M. O'Connell, and L. P. Lounibos. 2006. Habitat segregation of mosquito arbovirus vectors in south Florida. *Journal of Medical Entomology* 43 (6):1134-1141.

Roberts, R. S., and I. M. Foppa. 2006. Prediction of equine risk of West Nile Virus infection based on dead bird surveillance. *Vector-Borne and Zoonotic Diseases* 6 (1):1-6.

Rogers, D., and S. Randolph. 2000. The global spread of malaria in a future, warmer world. *Science* 289:1763 - 1766.

Ronco, A. L. 1999. Use of artificial neural networks in modeling associations of discriminant factors: towards an intelligent selective breast cancer screening. *Artificial Intelligence in Medicine* 16 (3):299-309.

Ruiz, M. O., C. Tedesco, T. J. McTighe, C. Austin, and U. Kitron. 2004. Environmental and social determinants of human risk during a West Nile virus outbreak in the greater Chicago area, 2002. *International Journal of Health Geographics* 3:8:2 - 11.

Ruiz, M. O., E. D. Walker, E. S. Foster, L. D. Haramis, and U. D. Kitron. 2007. Association of west Nile virus illness and urban landscapes in Chicago and Detroit. *International Journal of Health Geographics* 6 (10):1 - 11.

Saelens, B., J. Sallis, J. Black, and D. Chen. 2003. Neighborhood-based difference in physical acitivity: an environment scale evaluation. *American Journal of Public*

Health 93:1552 - 1558.

Sattler, M., D. Mtasiwa, M. Kiama, Z. Premji, M. Tanner, G. Killeen, and C. Lengeler. 2005. Habitat characterization and spatial distribution of Anopheles sp. mosquito larvae in Dar es Salaam. *Malaria Journal* 4 (4).

Savage, H., and B. Miller. 1995. House mosquitoes of the USA. *Florida Mosquito Association* 6 (2):8 - 9.

Savage H., and B. Miller. 1995. House mosquitoes of the USA. *Florida Mosquito Association* 6 (2):8 - 9.

Savage, H. M., M. Anderson, E. Gordon, L. McMillen, L. Colton, D. Charnetzky, M. Delorey, S. Aspen, K. Burkhalter, B. J. Biggerstaff, and M. Godsey. 2006. Oviposition activity patterns and West Nile virus infection rates for members of the Culex pipiens complex at different habitat types within the hybrid zone, Shelby County, TN, 2002 (Diptera : Culicidae). *Journal of Medical Entomology* 43 (6):1227-1238.

Schafer, M. L., J. O. Lundstorm, M. Pferrer, and E. Lundkvist. 2004. Biological diversity verses risk for mosquito nuisance and disease transmission in constructed wetlands in southern Sweden. *Medical Veterinary and Entomology* 18:256 - 267.

Schikowski, T., D. Sugiri, V. Reimann, B. Pesch, U. Ranft, and U. Kramer. 2008. Contribution of smoking and air pollution exposure in urban areas to social differences in respiratory health. *Bmc Public Health* 8:179.

Schinka, J. A., W. F. Velicer, and I. B. Weiner. 2003. *Research methods in psychology*: Wiley.

Shaman, J. 2007. Amplification due to spatial clustering in an individual-based model of mosquito-avian arbovirus transmission. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 101 (5):469-483.

Shaman, J., J. F. Day, and M. Stieglitz. 2005. Drought-Induced amplification and Epidemic Transmission of West Nile virus in Southern Florida. *Journal of Medical Entomology* 42 (2):134 - 141.

Smith, D., J. Dushoff, and F. McKenzie. 2004. The risk of a mosquito-borne infection in a heterogeneous environment. *Plos Biol* 2:e368.

Steele, K. W., M. J. Linn, and R. J. Schoepp. 2000. Pathology of fatal West Nile virus infections in native and exotic birds during the 1999 outbreak in New York City, New York. *Vetenarary Pathology* 37:208 - 224.

Stockwell, P. J., N. Wessell, D. R. Reed, T. A. Kronenwetter-Koepel, K. D. Reed, T. R. Turchi, and J. K. Meece. 2006. A field evaluation of four larval mosquito control methods in urban catch basins. *Journal of the American Mosquito Control Association* 22 (4):666-671.

Tachiiri, K., B. Klinkenberg, S. Mak, and J. Kazmi. 2006. Predicting outbreaks: a spatial risk assessment of West Nile virus in British Columbia. *International Journal of Health Geographics* 5 (1):21.

Theophilides, C. N., S. C. Ahearn, E. S. Binkowski, W. S. Paul, and K. Gibbs. 2006. First evidence of West Nile virus amplification and relationship to human infections. *International Journal of Geographical Information Science* 20 (1):103 - 115.

Theophilides, C. N., S. C. Ahearn, S. Grady, and M. Merlino. 2003. Identifying West Nile virus risk areas: The dynamic continuous-area space-time system. *American Journal of Epidemiology* 157 (9):843-854.

USDA, U. S. D. o. A. 1977. Soil Survey of Anoka County, Minnesota. In *Minnesota Online Soil Survey Manuscripts*: National Resource Conservation Services.

Vaidyanathan, R., and T. W. Scott. 2006. Seasonal variation in susceptibility to West Nile virus infection in Culex pipiens pipiens (L.) (Diptera : Culicidae) from San Joaquin County, California. *Journal of Vector Ecology* 31 (2):423-425.

Warner, R. D., R. C. Kimbrough, J. R. Pierce, T. Ward, and L. P. Martinelli. 2006. Human West Nile virus neuroinvasive disease in Texas, 2003 epidemic: Regional differences. *Annals of Epidemiology* 16 (10):749-755.

Watson, J. M., H. L. Logan, and S. L. Tomar. 2008. The influence of active coping and perceived stress on health disparities in a multi-ethnic low income sample. *Bmc Public Health* 8:41.

Williot, E. 2004. Restoring nature, without mosquitoes? . *Restoring Ecology* 12:147 - 153.

X-Pro. X-Pro Tools for ArcGIS Extensions. http://www.xtoolspro.com/.

Yiannakoulias, N. W., D. P. Schopflocher, and L. W. Svenson. 2006. Modelling geographic variations in West Nile virus. *Canadian Journal of Public Health-Revue Canadienne De Sante Publique* 97 (5):374-378.

Yiannakoulias, N. W., and L. W. Svenson. 2007. West Nile Virus: Strategies for Predicting Municipal-Level Infection. *Annals of New York Academy of Sciences* 1102:135 - 148.

Zou, L., S. Miller, and E. Schmidtman. 2007. A GIS tool to estimate West Nile virus risk based on a degree-day model. *Environmental Monitoring and Assessment* 126:413 - 420.

Zou, L., S. N. Miller, and E. T. Schmidtmann. 2006. Mosquito larval habitat mapping using remote sensing and GIS: Implications of coalbed methane development and West Nile virus. *Journal of Medical Entomology* 43 (5):1034-1041.