This is to certify that the
thesis entitled

SPATIAL TRENDS OF WEST NILE VIRUS IN DETROIT, MICHIGAN 2002

presented by

Kevin Patrick McKnight

has been accepted towards fulfillment
of the requirements for the

Master of Arts degree in Geography

Major Professor's Signature

14 DEC 06

Date

SPATIAL TRENDS OF WEST NILE VIRUS IN DETROIT, MICHIGAN 2002

By

Kevin Patrick McKnight

A THESIS

Submitted to
Michigan State University
In partial fulfillment of the requirements
for the degree of

MASTER OF ARTS

Department of Geography

2006

ABSTRACT

SPATIAL TRENDS OF WEST NILE VIRUS IN DETROIT, MICHIGAN 2002

By

Kevin Patrick McKnight

West Nile Virus is vector-borne flavivirus that affects mainly birds, horses, and humans. The disease emerged in the United States in 1999 and by 2001 had reached Michigan. Currently, the virus has been reported in all 48 contiguous states. In clinical human cases, the most common symptoms are fever, weakness, nausea, headache, and changes in mental state. The crow is the most common wildlife host in the life cycle of the virus. The state of Michigan, through the Michigan Department of Community Health, collected the spatial locations of dead birds (Corvidae). Statewide, during 2002, there were over 8,000 reports. The large number of samples made spatial and temporal hotspot detection possible. However, the volunteer reporting method produced a dataset with a direct correlation between the numbers and locations of the dead birds and population density. Therefore, accurately identifying hotspots remains a challenge. Extensive cleaning was required to insure the data points were spatially accurate. The dataset was then modeled using Ripley's K, Moran's I, Oden's I(pop), Local Moran's I, Kulldorff's Spatial Scan, and the Geographic Analysis Machine. These statistical models identified overall spatial structure and local clustering in the dataset. Identification of hotspots was confounded by limited information about the collection procedures, data availability, and the limitations of each method.

Dedicated to my parents:
This would not have been possible
without your support.

Special Thanks to:
Edward F. Hartwick and Beth M. Clute

I would also like to thank my advisor, Dr. Joseph Messina, and my committee members, Drs. Bruce Pigozzi and Ashton Shortridge. Your advice throughout this research was invaluable.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Ripley's K Maps

Anselin's Local Indicator of Spatial Autocorrelation Maps

Kulldorff Scan Method Maps

Geographic Analysis Machine Maps

Images in this document presented in color.

1 Introduction

1.1 Introduction

This research was designed to examine the geographic extent that disease data collected from the public can be used in spatial and temporal cluster analyses. The term "geographic" refers to the Earth's surface and near-surface, and defines the subject matter of this thesis; however other terms have similar meaning. The term "spatial" is used frequently throughout this document, almost always with the same meaning as "geographic" (Longley, Goodchild et al. 2001). The underlying structure patters of the disease data will be analyzed using spatial statistics. The statistical models progress from global models that identify general spatial structure, to local models that locate disease clusters. The introductory chapter provides an overview of medical geography, West Nile Virus, and spatial statistics. This section followed by a literature review focusing on the history, virology, and pathology of West Nile Virus. The literature review also discusses the spatial statistical models used in the research.

Understanding the factors that allow, or more importantly, cause a disease to spread, greatly affects the manner in which the public health officials respond. By examining the spatial locations of disease cases, as well as the environmental and social milieu, patterns and commonalities may emerge among the individual events. These patterns often reflect underlying environmental influences. Medical geography attempts to identify the relationship between diseases and their spatial context.

Medical geography uses the concepts and techniques of the discipline of geography to investigate health-related topics (Meade and Earickson 2000). The ability to combine

Geographic Information Systems (GIS) with medical diagnoses allows for the novel assessment of a disease over a geographic region. By using location as a factor in disease research, spatial statistics enable researchers to examine the quantitative and qualitative patterns associated with the disease. Many disease characteristics, for example risk and extent, can be identified using these types of geospatial analyses. Further, clusters and point sources may be located. Also, the direction and intensity of a spreading virus epidemic can be estimated. Knowing how a virus interacts with the environment is critical in controlling the spread of the disease.

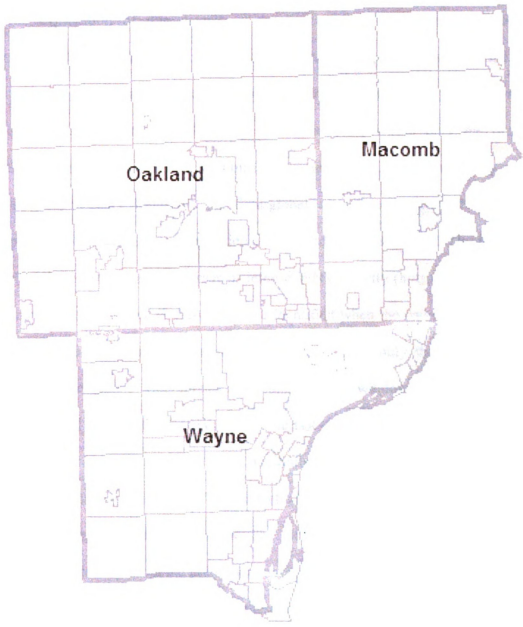
Spatial statistics encompass a collection of tools that are used to analyze patterns and trends in data across a map. These spatial techniques are used to find or describe the extent of clustering or autocorrelation across a given area. For the purpose of this study, structure or autocorrelation describes any deviation from complete spatial randomness, such as regularity, clustering, or a first order trend. Global models only examine if structure is present and not if the data points are clustered. Local models explore the spatial dependence of deviations in attribute values from their mean; that is, the second order properties (Bailey and Gatrell 1995). If a second order trend is found to be present, then the point pattern is referred to as being clustered. These methods can be traced back to the famous case of John Snow's evaluation of the 1854 cholera epidemic in London, where the pattern of disease cases was observed and analyzed to determine if the pattern was of a global or local nature. Techniques such as cluster detection have grown in demand with the increasing availability of high accuracy spatial datasets. Identifying the locations of data points is often done for individual cases by geocoding (process of

assigning geographic identifiers to data records, such as street addresses) the address of the event (Lawson and Kleinman 2005), and is used for the dataset in this research.

West Nile Virus became endemic (a disease native to a particular people or specific region) to the Western Hemisphere in 1999. It surfaced in New York City causing 67 human cases and seven deaths (Craven and Roehrig 2001). During the next few years the disease spread across the United States through a bird-mosquito-bird cycle. The birds act as the reservoir for the virus and mosquitoes spread the virus throughout susceptible populations. In the beginning of the 2000 transmission cycle it was confirmed that at least 12 species of mosquito could spread the disease (Craven and Roehrig 2001). Currently 43 species of mosquito have tested positive for West Nile Virus (WNV) (Marra, Griffing et al. 2004). The *Culex* species of mosquito is the most common bridge vector transferring the virus from birds to humans. Since 1999, the disease had infected at least 400,000 humans, and killed countless birds, mammals, and reptiles (Dodd 2003; Marra, Griffing et al. 2004).

The counties presented in this thesis are, by design, limited to urban or populated areas. The Greater Detroit Area is representative of urban areas in Michigan. Wayne, Oakland, and Macomb counties of the Greater Metropolitan Detroit Area were selected for this research (Fig. 1), because they contained the most dead bird records and the greatest at risk population. With a fairly consistent population density and widespread urbanization the Metro Detroit Area provides the largest uniform study area available where many events were reported. This region of the state manages its own assets through the South East Michigan Commission of Governments (SEMCOG). The road framework data used to georeference the WNV bird data are some of the most accurate in

the state. Dense population, high number of dead bird records, highly urban landcover, and accurate GIS information combine to make the Greater Detroit Area the overall best site for this research in the state of Michigan.



Map of study area, urban areas shown in gray.
Fig. 1

The primary objectives of this research are to explore geographic epidemiology of West Nile Virus in the Detroit Metropolitan Area in 2002, though the analysis of the spatial point patterns developed from the citizen reported dead bird data, and to compare and contrast repeatable methods that permit the identification of disease clusters.

The exploration of this dataset is designed to provide information about the distribution of WNV. The three global models will demonstrate that WNV in the Detroit area presents global spatial autocorrelation across the study area. The progression of the global models will attempt to identify spatial structure within only the dead bird locations; first by examining the dead bird data as individual level data (Ripley's K) and then as grouped data (Moran's I). The final global test will examine if autocorrelation is still present when the human population is included (Oden's I(pop)). The local spatial models will be used to identify areas of intense WNV activity (hotspots). The first two local models will examine if clusters can be identified when the dataset is aggregated to grid cells. The first method examines the relationship of only the neighboring cells (LISA) and the second incorporates a circular moving window that increases until a cluster is found (Kulldorff's Spatial Scan). The last local method will examine the same dataset but as individual points using a circular moving window at various scales (Geographic Analysis Machine). This research also explores the temporal distribution of the virus through the infection season. The diffusion of the virus can be mapped in direction and intensity using the point locations along with the date the birds were found.

Given the process used to report these data, certain assumptions needed to be addressed before statistical modeling was possible. The first assumption was that the dataset contains 100% of the dead birds. This was necessary because the statistical

methods operate under the assumption that the dataset is comprised of an entire population. By assuming that the dataset contains 100% of the dead birds, it must also be assumed that all of the dead birds in the dataset were WNV positive. The infectivity of the birds must be assumed due to the limitations used in testing the birds. After the first positive test, the subsequent birds found in the county were considered positive. These assumptions introduced some error into the dataset. However, the extent of the error cannot be quantified due to the collection process and insufficient data on the entire crow population.

The remainder of this thesis examines a methodology for preparing and modeling a spatial dataset. The results of the global and local models are then examined for spatial trends and disease hotspots. In closing, final conclusions and recommendations for further research are discussed.

1.2 Literature Review

To understand the geographic aspects of West Nile Virus there are three main topics that need to be examined. The first section reviews the origins of the disease and how it spread to the United States. This is required to track the path of the virus and examine possible ways to stop its spread. The second section is on the disease itself. It discusses medical descriptions of the virus, disease rates, human symptoms, and treatment. The final section examines the transmission of the disease, the mosquito (vector), and the crow (reservoir).

West Nile Virus (WNV) was first identified in 1937 in the West Nile region of Uganda (Petersen and Marfin 2002). Since the discovery of the virus, there have been sporadic outbreaks across Europe, Africa, and Asia. The first large-scale outbreak occurred in Israel from 1950 to 1957. Hundreds of clinical cases were reported. In 1974, an epidemic of WNV spread across South Africa producing tens of thousands of infections. With the exception of a few cases in France in 1962, the Middle East and Africa were the known extent of the virus.

It was not until 1996 when a large-scale outbreak occurred in Europe that the disease moved out of Africa. The mid-August outbreak, around Bucharest in Southeastern Romania, saw 835 patients admitted to hospitals showing signs of central-nervous-system infections. Of these cases, 767 met the clinical definition of WNV, and 441 of those had the proper blood and cerebrospinal fluid samples necessary for confirmation of the virus. From this sample, there were 352 confirmed cases, but there were also 41 patients who did not show symptoms who had positive tests.

The World Health Organization confirmed 393 infections. Among the clinical cases, the diagnoses were meningitis (40%), meningoencephalitis (44%), and encephalitis (16%) (Tsai, Popovici et al. 1998). The illness progressed to coma in 13% of cases, and there were 17 fatalities out of all patients over 50 years of age. Age is the main contributing factor in the fatality rate associated with WNV (Tsai, Popovici et al. 1998). From birth to age 50 the case fatality rate during the Romania outbreak was zero, 3.4% for persons age 50 – 59, 4.3% for ages 60 – 69, and 14.7% in those over 70 years. The disease attack rate in Bucharest was reported to be 12.4 per 100,000 people.

The Tsai study (1998) identified the *Culex pipiens* as the vector responsible for bridging the disease from birds. However they suggested that infected birds from Africa or the Middle East migrated into the area, and then the *Culex modestus* species of mosquito transmitted the virus throughout the birds in the area. As the virus spread through the bird population by *Culex modestus,* the *Culex pipens* acted as the bridge vector spreading the virus from the birds to the human population (Tsai, Popovici et al. 1998).

Until 1999 the virus was still limited to Africa, the Middle East, Western and Central Asia, India, Europe and Australia (where the virus is known as *Kunjin virus)* (Anderson, Vossbrinck et al. 2001). The disease emerged in the United States in 1999, in New York City. These were the first recognized cases in the Western Hemisphere. There were 59 patients from New York (almost entirely from Queens) diagnosed with advanced clinical symptoms of the disease. After the initial outbreak, the virus quickly spread across the United States through the South and Midwest. It was spread through the bird migrations across the country establishing endemism in humans, mosquitoes, mammals (horses) and birds. As of 2002, there were positive reports of the virus in all states except Arizona, Utah, Nevada, Oregon, Alaska, and Hawaii (Huhn, Sejvar et al. 2003). Currently (2006), there have been confirmed cases of WNV in all 48 contiguous states, and since 1999 it has become an endemic disease to North America. The spread of this virus caused the largest human arboviral encephalitis epidemic in United States history (Huhn, Sejvar et al. 2003). From June 10 to December 31, 2002, 4,156 cases of WNV infection (including 284 deaths) were reported across 39 states and the District of Columbia (Huhn, Sejvar et al. 2003). This huge increase in positive infections was probably due to the expanded

knowledge of the virus among those in the medical community, and to public recognition that symptoms required medical attention (Huhn, Sejvar et al. 2003).

The geographic diffusion of WNV across the United States in 2002 was similar to a large (2,131 cases) epidemic of St. Louis encephalitis in 1975. Both viruses concentrated in the Ohio and Mississippi river valleys. It was recognized that there was a geographic distribution of the time of infection for human cases for WNV (Huhn, Sejvar et al. 2003). In the southern states, infection occurred from the summer months until December, which is beyond the characteristic arboviral disease-transmission season of June to October (Huhn, Sejvar et al. 2003). In the northern states, the infections started later in the summer and ended earlier in the fall. The South had a longer season with a lower transmission peak, where as the North had a slightly shorter transmission season but the intensity of the peak weeks was two to five times higher. This was most likely due to the seasonally cooler weather, causing a shorter but more intense breeding season for the mosquitoes. (Huhn, Sejvar et al. 2003).

WNV is a vector-borne RNA flavivirus that affects mainly birds, horses, and humans (Anderson, Vossbrinck et al. 2001). However, all mammals and some reptiles are at risk. The virus is maintained in the United States in an enzootic, mosquito-bird-mosquito cycle. The most common mosquito vector known to spread the virus is the *Culex* species, however it has been isolated or detected in greater than 20 species of mosquitoes in the eastern United States (Anderson, Vossbrinck et al. 2001). These mosquitoes develop in the early spring from aquatic stages and begin to infect birds. The birds (mainly corvids) then become the reservoir for the disease; subsequent mosquito blood meals complete the cycle (Craven and Roehrig 2001).

Risk associated with this virus is difficult to assess, as there are many risk factors such as, age, previous illnesses, and viral dosage. Since most of the cases never develop recognizable symptoms, those sub-clinical cases go unreported. Thus, the overall risk to the population cannot be accurately calculated. However the risks to those patients who seek medical attention, or are hospitalized can be calculated. From patients hospitalized with this disease the case fatality rate ranged from 4% in Romania (1996) to 12% in New York (1999), and 14% in Israel (2000) (Petersen and Marfin 2002). The New York case fatality rate remained constant throughout the United States in 2000 and 2001. Age is the dominant risk factor associated with the disease. In New York, persons 75 and older were nine times more likely to die than younger persons (Petersen and Marfin 2002). Not much is known about the long-term morbidity associated with post-hospitalization for WNV infection but may be quite substantial (Petersen and Marfin 2002). Of the people who were hospitalized in New York and New Jersey in 2000 half did not fully recover by the time of discharge. Follow-ups on the patients from the 1999 outbreak found persistent symptoms such as fatigue (67%), memory loss (50%), muscle weakness (44%), and depression (38%) (Petersen and Marfin 2002).

Only about one in 150 infected humans will develop recognizable symptoms. The symptomatic illness has an incubation period of two to 15 days. A study of the 59 hospitalized patients from the New York City outbreak found the most common symptoms were fever (90%), weakness (56%), nausea and vomiting (51-53%), headache (47%), and changes in mental state. Advanced cases may develop into meningitis, encephalitis, flaccid paralysis, coma, and death (Huhn, Sejvar et al. 2003).

The treatment for WNV is supportive, and there are currently no licensed medications for the prevention of the disease. In order to prevent infection of WNV on a personal level, the recommended measure is protective clothing during the high-risk season (April to October). The state and local level prevention plan most often suggested in the literature was to identify the vector species breeding sites and use aggressive surveillance to determine the risk posed by the vector (Enserink 2000; Huhn, Sejvar et al. 2003).

Mosquitoes may live up to several weeks, and the females are capable of laying multiple batches of eggs in their short lifespan. The females of some mosquito species require vertebrate blood to produce each batch of eggs; however this is not always a necessity, blood from any animal would provide the proper proteins. Many mosquitoes have a specific preference (e.g., birds, mammals, or reptiles) for their blood meal. The time between the blood meal from an infectious host and when the mosquito is able to transmit the virus varies by species and environment. It is possible for the cycle to occur in as little as four to five days (at high ambient temperature around 26-30°C) and the testing showed mosquitoes were still able to transmit the disease at 32 days (Dohm, O'Guinn et al. 2002).

Mosquito species also vary in their ability to transmit WNV. Among laboratory-tested species, it was found that different "doses" of WNV were required to cause infection with different mosquito species. It was also found that WNV does not reach the salivary gland in all mosquito species, so not all infected mosquitoes are capable of spreading the virus through blood feedings (Marra, Griffing et al. 2004).

The main vector that is associated with the avian spread of WNV is the *Culex pipiens* species of mosquitoes. This species breeds in standing water, urban drains and catch

basins (Allen and Shellito 2004). Normally, the *Culex* species will breed from April to October in wet, richly organic areas such as swamps, fresh water rivers, and on poorly drained agricultural land (Allen and Shellito 2004). These areas, as opposed to urban microenvironments (e.g. old tires), allow for the greatest breeding potential of the vector. *Culex pipiens* species transmit WNV through a bird-mosquito-bird cycle. There are many bird species susceptible to WNV and act as hosts for the virus. The best recognized host for this cycle is the American Crow (*Corvus branchyrhynchus)*. This species has a high case fatality rate, and is recognized as the main reservoir of the virus (Komar 2000; Eidson, Kramer et al. 2001). The crow presents a high rate of infection and morbidity from this disease with death occurring four to eight days after infection, following exposure (McLean, Ubico et al. 2001; Komar, Langevin et al. 2003). During the outbreak in New York City in 1999 nearly 89% of the dead birds that were collected and had laboratory confirmed WNV infections were American crows (Komar 2000).

In order to examine the dataset of dead birds for structure, spatial point pattern analysis will be used. Spatial point pattern analysis became common in geography in the late 1950s and early 1960s. At that time, the desire to examine spatial relationships in datasets was at the forefront of the discipline. The techniques were adapted from the literature relating to plant ecology. These methods of analyses were slow to gain popularity due to the absence of suitable software tools. The first programs published generated textual and numerical output, but did not have graphing and map capabilities (Gatrell, Bailey et al. 1996). The advances in geographic information systems provided the proper tools for these type of data to be analyzed and displayed more efficiently (Gatrell, Bailey et al. 1996).

There are many methods for examining spatial point patterns such as "nearest-neighbor" and "quadrat" analyses. However these two methods look for structure within a dataset of points and do not allow for the inclusion of underlying causes, such as population. The exclusion of population data is a problem with the study of disease epidemiology in that the pattern of most diseases is directly related to the density of a population. As such, spatial variation in patterns of disease is often explored with techniques of cluster detection analysis (Klassen, Kulldorff et al. 2005). Detecting clusters in spatial data can be done many ways, for the purposes of this study three cluster detection methods will be used, Global, Local, and Temporal.

Global cluster detection methods are used to search for spatial patterns across the entire study area. Global methods determine whether structure exists within the dataset, and if the structure is unlikely to have arisen from chance. The statistics modeled for this method are, Ripley's K, Moran's I, and Oden's I(pop). Local methods look for clusters in a particular area. They examine the proximity of cases and determine if they are closer than would be expected by chance. The models that were preformed for this method are Anselin's Local Moran, Kulldorff's Spatial Scan Test, and the Geographic Analysis Machine (GAM). Temporal tests examine the data over time. Using the results from the previous models and dividing the data into smaller divisions of time, variations in the intensity of the virus may indicate peaks during the transmission season.

2 Methodology

2.1 Data

The Michigan Department of Community Health (MDCH) collected information on dead bird sightings via a toll-free hotline and a web report system. The hotline was a message service where citizens could leave information on the species of bird, the street address of where the bird was observed, the date of the observation, whether the bird had been collected, and a contact number if the person was willing to have the bird collected for West Nile Virus testing. The original statewide dataset contains 8,249 dead bird records. From the original database only the corvids (crows and blue jays) were selected for geocoding. The dead bird sightings used for this research were collected from April through December of 2002.

2.1.1 Dead Crow Data

The dead bird database used the address information of where the bird was found to place the location on the map. Most records also included township and county data. This assisted in confirming the proper location when address matching. Address matching is a process that uses road addresses to append points to an estimated spatial location along a road segment using a GIS road layer. This was done, using ESRI GIS software, for the entire state. This produced a new database (shapefile format) that contained geocoded points representing the locations where the dead birds were found.

Due to the nature of the data collection process, citizen reported and not systematically constructed, there were fields in the dataset that contained no data or

improperly entered data. These points failed to meet location accuracy standards. In order to reduce error to a minimum the dataset was first clipped to the spatial extent of the study area (Oakland, Macomb, and Wayne counties). From the original 8,249 statewide records the three county study area dataset contained 2,533 bird records. All records were individually checked for a positive road match; also crossroad information was verified, if present. The county field was left blank for some cases in the original database. If the road information were correct, that point would be placed in the county to which it was addressed. In addition, if the county name was entered and incorrect, the point was removed.

With the dataset cleaned of unreliable records, 2,497 dead bird points remained, coincidently, only crows. With the file "cleaned" the data points were attributed with spatial boundary information. Using a tagging function, township, city, and block group information was appended to each data point. The data preparation process was completed using TransCAD (Caliper 2006). At this point, the data set was considered complete, closed for modification, and suitable for statistical analyses. As Table 1 illustrates, most of dead birds were found in Wayne County, with Oakland and Macomb Counties having comparable ratios of dead birds to people.

Records by County

County	Dead Bird Records	Population	Ratio
Oakland	887	1,194,156	.00074278
Macomb	502	787,625	.00063736
Wayne	1,108	2,061,162	.00053756
TOTAL	2,497	4,042,943	.00061762

Table 1.

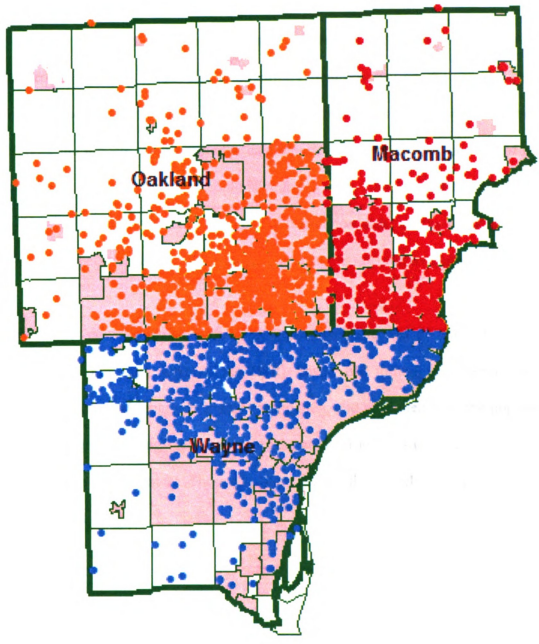
The data were then divided by month and within the month over the course of the transmission season (Table 2).

Records by Month

Month	Records
April	**2**
May	**32**
April & May	34
June	**59**
June 1-15	34
June 16-30	25
July	**396**
July 1-10	20
July 11-20	108
July 21-31	268
August	**1,826**
August 1-10	875
August 11-20	461
August 21-31	490
September	**166**
October	**16**
TOTAL	**2,497**

Table 2.

The distribution of dead bird records can be seen in Figure 2.

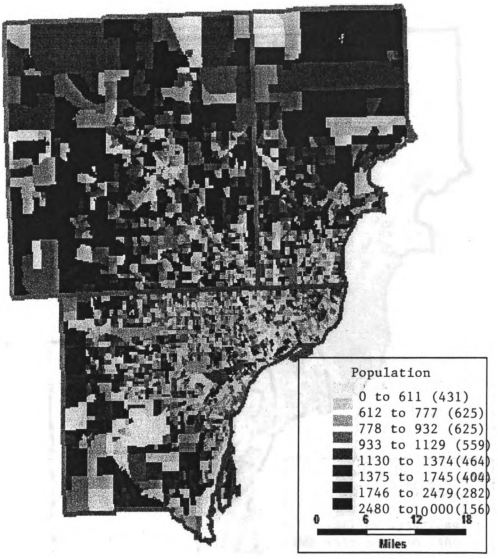


Dead bird distribution by county.
Fig. 2

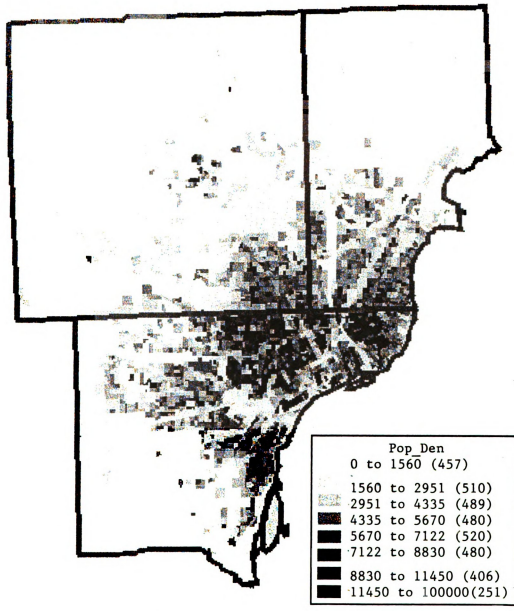
2.1.2 Census Data

Given the method of collection, it was likely that the underlying human population influenced the bird data simply due to opportunity and observation, the observation bias attributed to population had to be accounted for in the models. By incorporating human population, the models can examine the data objectively and compare the ratio of found birds to the number of people who could have found them. Population data were obtained from the 2000 United States Census. For the population data, census block group data were selected, because it was the finest spatial resolution dataset available for the study area.

The dataset downloaded from the U.S. Census web site was for the entire state of Michigan. The raw data were in the ESRI Shape file format, as centroid points, containing latitude and longitude spatial information, and attributed with the population for each block group. Using TransCAD, only the points from the study area were selected and edited to the proper extent. These data were used as the underlying population for all statistical models.



Human population by block group.
Fig. 3



Population density by block group.
Fig. 4

2.2 Global Statistics

2.2.1 Ripley's K

Ripley's K-function was used to analyze the spatial pattern of point data by considering its density within a set of distances, and to detect global spatial structures in individual-level data by comparing the observed proximity of cases with the pattern generated by a homogenous Poisson process. Ripley's K-function describes the extent to which there is spatial dependence in the arrangement of events. The K-function is estimated for the observed data, and then compared to an arrangement of events showing complete spatial randomness (CSR). The CSR pattern used to model the dataset is created by a homogenous Poisson process. It can also be compared using a Monte Carlo randomization of the data (Bailey and Gatrell 1995).

The expected number of other events within a fixed distance (h) of one event is $\lambda K(h)$, where λ is the intensity, or mean number of events per unit area. The variable h is the radius of a circle from each event used to examine the point pattern (Fig. 5).

K(h) or *Khat* can be estimated by the following formula (Bailey and Gatrell 1995)

$$\hat{\mathrm{K}}(h) = \frac{R}{n^2} \sum_{i=1}^{n} \sum_{j=1, i \neq j}^{n} \frac{I_h(d_{ij})}{w_{ij}} \qquad \textbf{(1)}$$

Here R is the area of the region of interest, n is the total number of events in region R, d_{ij} is the distance between the i^{th} and j^{th} events, and $I_h(d_{ij})$ is the indicator function which is

1 if $d_{ij} \leq h$ and 0 otherwise. The function sums the number of events within distance h of each location in the dataset (each i). Typically w_{ij} is the conditional probability that points around event i will be in the study area. The Ripley's K model calculates the weight as a proportion of the circle's area that falls within the study area. The weight for a circle falling completely within the study area would be 1, and if half of the circle's area is outside of the study area the weight would be 0.5. The event count in that area is essentially doubled to account for the missing half of the circle (Fig. 6). If *Khat* is estimated for a number of distances, one can examine the spatial dependence of points relative to distance. In addition, the estimate of *Khat* can be converted/normalized to a test statistic *Lhat* that permits the significance of the deviation of an observed value of *Khat* from its expected value (under the assumption of randomness according to a Poisson distribution) to be tested (Fortin 1999); if *Lhat* is significantly different from zero then the distribution of points is not random. The general form of *Lhat is:*

$$Lhat_h = (Khat_h/\pi)^{0.5} - h \quad \textbf{(2)}$$

Ripley's K results are reported as *Lhat* in the graph outputs. The null hypothesis for Ripley's K is the distribution of events is a spatial Poisson point process (complete spatial randomness). The alternative hypothesis is that the dataset shows structure at some scales.

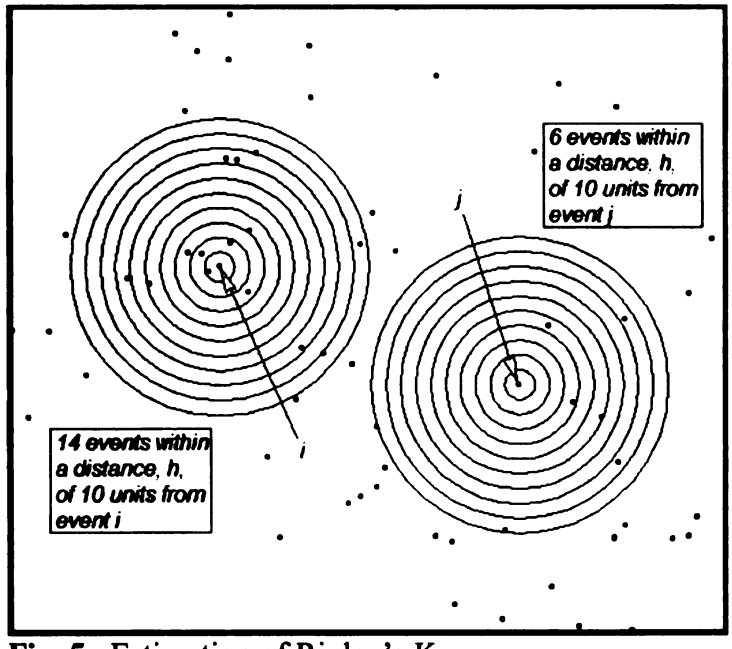


Fig. 5 Estimation of Ripley's K

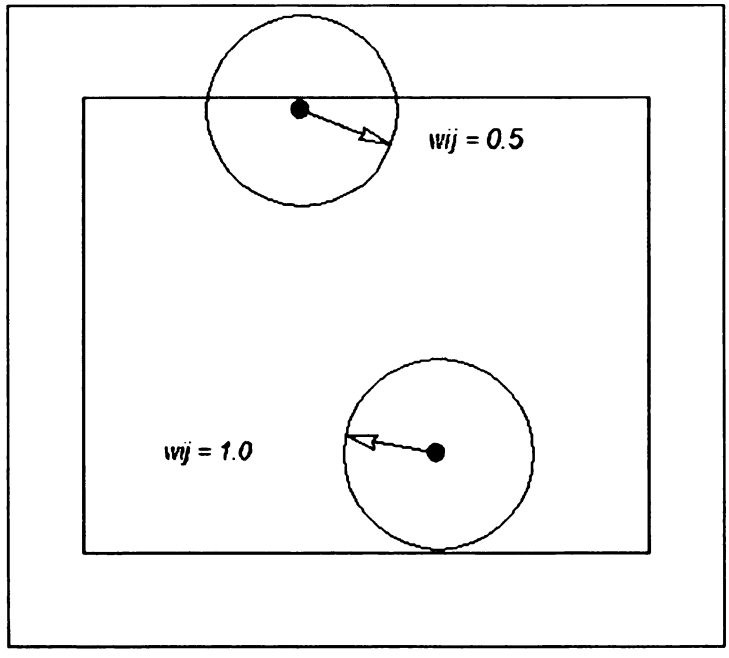


Fig. 6 Estimating the weights for Ripley's K

2.2.2 Moran's I

Moran's I is a weighted correlation coefficient used to detect departures from spatial randomness (Moran 1950). Departures from randomness indicate patterns such as

clusters, but may also identify geographic trends. Moran's I examines global spatial autocorrelation in group-level data. Positive spatial autocorrelation means that nearby areas have similar autocorrelation rates, this would indicate spatial structure within the dataset. Nearby areas have similar autocorrelation rates when the populations and exposures are alike. When rates are similar, Moran's I will be large and positive. A Moran's I value close to 1 is indicative of positive autocorrelation or a clustered spatial structure. When rates in nearby areas are dissimilar, Moran's I will be negative. A negative Moran's I value indicates negative spatial autocorrelation, or regularity in the point pattern.

Moran's I (Moran 1950) is used to determine whether neighboring areas are more similar than would be expected under the null hypothesis. The null hypothesis for Moran's I, is that the disease rates are spatially independent and that the observed rates are assigned at random among locations. If I is close to 0 then the null hypothesis is accepted. The alternative hypothesis is that the disease rates are not spatially independent. The null hypothesis would be rejected if I were not 0.

Moran's I is defined as:

$$I = \frac{N \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} z_i z_j}{S_0 \sum_{i-1}^{N} z_i^2} \qquad \textbf{(3)}$$

Where, N equals the number of regions, w_{ij} is a weight representing the intensity of the connection between areas i and j, z_i is the rate in zone i centered around the mean rate (using $z_i = x_i - ave(x)$; x_i is the rate in zone i); and S_O is the sum of the weights.

$$S_0 = \sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}, i \neq j$$

(4)

The expectation of I under the null hypothesis is:

$$E(I) = \frac{-1}{(N-1)}$$

(5)

The expectation gets closer to 0 as N increases. The variance of I is determined under two null hypotheses or assumptions: Normality (denoted **N**) or randomization (denoted **R**). Under assumption **N** the rates are sampled from a mean-zero population whose distribution is normal. Under the **R** assumption the rates are random samples from a population whose distribution is assumed to not follow a normal distribution. Assumption **N** is useful when the observations are thought to follow a normal distribution. Assumption **R** is less restrictive and, since we often don't know the theoretical distribution, is appropriate for disease rates.

The variance under assumption **N** is:

$$Var_N(I) = \frac{1}{(N-1)(N+1)S_0^2}\left(N^2 S_1 - NS_2 + 3S_0^2\right) - E(I)^2$$

(6)

The variance under the assumption **R** is:

$$Var_R(I) = \frac{N\left|\left(N^2 - 3N + 3\right)S_1 - NS + 3S_0^2\right| - b_2\left|\left(N^2 - N\right)S_1 - 2NS_2 + 6S_0^2\right|}{(N-1)^{(3)}S_0^2} - E(I)^2 \tag{7}$$

Where, a falling factorial is written $s^{(b)} = s(s-1)...(s-b+1)$,

and where

$$\begin{aligned} b_2 &= m_4 / m_2^2 \\ m_4 &= 1/N\sum_{i=1}^{N} z_i^4 \\ m_2 &= 1/N\sum_{i=1}^{N} z_i^2 \\ S_1 &= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(w_{ij} + w_{ji}\right)^2 \\ S_2 &= \sum\left(w_{i\bullet} + w_{\bullet i}\right)^2 \end{aligned} \tag{8}$$

Significance for the Moran's I model is evaluated under assumptions **R** and **N**, and by Monte Carlo simulations. For assumptions **R** and **N** the model calculates two z-scores as:

$$z_N = \frac{I - E(I)}{\sqrt{Var_N(I)}} \tag{9}$$

and

$$z_R = \frac{I - E(I)}{\sqrt{Var_R(I)}} \tag{10}$$

These z-scores express the difference between the observed and expected value of *I* in standard deviation units. The distribution of the z-scores is approximately normal with a mean of 0 and a variance of 1.0. The Moran's I model reports a two-tailed P-value because spatial pattern is of interest both when Moran's *I* is positive (rates in connected areas are similar) or negative (rates in connected areas are dissimilar).

2.2.3 Oden's I(pop)

One major issue with Moran's I, is that it does not take into account underlying variation in a collinear variable population, which for the purposes of this study is important due to the nature of the data collection process. This problem was corrected with Oden's I(pop) (Oden 1995). Oden adapted Moran's I to examine population data to detect departures from spatial randomness. When population data are not used, large differences in population hinder Moran's I to accurately detect spatial autocorrelation and spatial randomness. Like Moran's I, Oden's I(pop) is a global level model for grouped data.

This model was adapted from Moran's I (Moran 1950) to take population data into account for computing the departure from randomness for spatial pattern data. The datasets for the population and disease events need to be in numeric totals for each unique polygon. When there is autocorrelation within a region or in adjacent regions the

I(pop) statistic will increase. Large values of Oden's I(pop) indicate positive spatial autocorrelation or similar rates in connected areas; small values indicate negative spatial autocorrelation or dissimilar rates in connected areas. The range of Oden's I(pop) depends on population size.

The statistic I(pop) is as follows:

$$Ipop = \frac{N^2 \sum_{i=1}^{m} \sum_{j=1}^{m} w_{ij}(e_i - d_i)(e_j - d_j) - N(1-2\bar{b}) \sum_{i=1}^{m} w_{ij} e_i - N\bar{b} \sum_{i=1}^{m} w_{ii} d_i}{S_0 \bar{b}(1-\bar{b})} \quad \textbf{(11)}$$

where:

$$e_i = ni/N \text{ and } d_i = xi/X \quad \textbf{(12)}$$

Here m represents the number of cells in the grid, and N is the total number of cases in all of the areas. For a given region, ni is the total number of cases in area i, and e_i is the proportion of cases in area i ($e_i = ni/N$). For the population counts, X is the total size of the risk population in all areas; xi is the size of the risk population in area i, and d_i is the proportion of the population in area i, $d_i = xi/X$. Also, $e_i - d_i$ is the difference between the proportion of cases in area i and the number of cases expected given the area's population size; and b is the average prevalence, $b = N/X$, $b2 = 1/b(1-b) - 3$. Finally, w_{ij} is a weight denoting the strength of connection between areas i and j, developed from neighbor information.

The equations for calculating S_0 and S_1 are:

$$S_0 = X^2 A - XB \tag{13}$$

$$S_1 = X^3 E - 4X^2 F + 4XD \tag{14}$$

For calculating these variables the following formulas are used:

$$A = \sum_{i=1}^{m}\sum_{j=1}^{m} d_i d_j w_{ij} \qquad E = \sum_{j=1}^{m} d_i \left[\sum_{j=1}^{m}\left(w_{ij} + w_{ji}\right)\right]^2$$

$$B = \sum_{i=1}^{m} d_i w_{ii} \qquad F = \sum_{j=1}^{m} d_i w_{ii} \sum_{j=1}^{m} d_j \left(w_{ij} + w_{ji}\right)$$

$$C = \sum_{i=1}^{m}\sum_{j=1}^{m} d_i d_j \left(w_{ij} + w_{ji}\right)^2 \qquad G = \sum_{i=1}^{m} e_i w_{ii}$$

$$D = \sum_{i=1}^{m} d_i w_{ii}^2 \qquad H = \sum_{i=1}^{m}\sum_{j=1}^{m} w_{ij}\left(e_i - d_i\right)\left(e_j + d_j\right)$$

(15)

Under the null hypothesis, which is that there is no autocorrelation, I(pop) approaches zero when the population is large. This value *E(Ipop)* is represented as:

$$E(Ipop) = \frac{-1}{(X-1)} \tag{16}$$

Oden's I(pop) calculates the variance in two ways. First, it is based on a random distribution, which is good for disease rates (Cliff and Ord 1981). Secondly, if the distribution is assumed to be normal the variance is approximated.

This variance under the null hypothesis is:

$$Var_R(Ipop) = \frac{X\left[(X^2 - 3X + 3)S_1 - XS_2 + 3S_0^2\right] - b_2\left[X^{(2)}S_1 - 2XS_2 + 6S_0^2\right]}{(X-1)^{(3)}S_0^2} - E(Ipop)^2 \quad \textbf{(17)}$$

Approximation of variance *(Var_A)*:

$$Var_A(Ipop) = \frac{2A^2 + \frac{C}{2} - E}{A^2 X^2} \quad \textbf{(18)}$$

For examining significance, Oden's I(pop) uses three methods: z-scores and variance, Monte Carlo randomization, and using multinomial randomization. These methods report similar P-values. For the purpose of this research only the Monte Carlo randomization assumption was used.

2.3 Local Statistics

2.3.1 Anselin's Local Moran (LISA)

The local Moran model (Anselin 1995) detects local spatial autocorrelation in group-level data. It is similar to Moran's I (Moran 1950), a model for global spatial autocorrelation, but the local Moran decomposes Moran's I into contributions for each location, termed LISAs, for Local Indicators of Spatial Association. These indicators

detect clusters of either similar or dissimilar disease frequency values around a given observation. The sum of LISAs for all observations are proportional to Moran's I, and are also an indicator of global pattern. Therefore one can make two interpretations of LISA statistics: as indicators of local spatial clusters and as a diagnostic for outliers in global spatial patterns.

The range of the LISA statistic is not as readily interpretable as the global Moran's I statistic, the values will depend on the number of observations in the dataset. Therefore the statistic should be evaluated such that positive values indicate like areas surrounded by like areas. Negative values indicate outliers in the dataset that do not constitute clustering. Values near zero indicate either the local area under consideration or the average of the immediate neighbors of the cell, or both, are near zero.

2.3.2 Kulldorff's Spatial Scan Statistic

Kulldorff's Scan method (Kulldorff and Nagarwalla 1995; Kulldorff 1997) can detect local space or space-time clusters in group-level data. The scan statistic uses a circular window to identify high concentrations of cases in space and time. The area is divided into spatial zones and in each zone a circular window increases in size until it reaches a set upper size limit. The Kulldorff scan statistic then compares a measure of whether the observed number of cases is unlikely for a window of that size, using reference values from throughout the study area. By searching for clusters without specifying their size or location, the method provides a model with no pre-selection bias. Kulldorff developed two models, a Poisson model and a Bernoulli model. For a small number of cases, the two models are similar. The Bernoulli model is best when studying disease using a

population with the disease and a control group without the disease, while the Poisson model better answers questions when using case and population-at-risk counts.

For the spatial scan, a circular window is moved systematically through the study area. The scan window starts at each location in the dataset. The spatial scan model uses the centroid of each polygon as the location for the window to start. The window expands to include the nearest region centroids. The maximum size of each window will not exceed 50% of the total population-at-risk size for the study period.

The hypotheses are evaluated with a maximum likelihood ratio test that examines whether the null or alternative model better fits the data (Kulldorff 1999). The scan statistic is the maximum likelihood ratio over all possible window sizes. Its P-value is obtained through multinomial Monte Carlo randomizations. If the null hypothesis is rejected, the spatial location and the extent of the cluster that caused the rejection were reported.

The likelihood ratio is:

$$\frac{L(Z)}{L_0} = \frac{\left(\frac{n_Z}{\mu(Z)}\right)^{n_Z}\left(\frac{N-n_Z}{N-\mu(Z)}\right)^{N-n_Z}}{\left(\frac{N}{\mu(A)}\right)^{N}} \qquad (17)$$

if $n_Z > \mu(Z)$, $1/L_0$ otherwise.

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

2.3.3 Geographic Analysis Machine (GAM)

The Geographic Analysis Machine (GAM) was developed in 1987 by Stan Openshaw at the Center for Computational Geographics at the University of Leeds (Openshaw, Charlton et al. 1988; Openshaw 1995). Originally, it was designed to identify disease clusters such as childhood leukemia (Openshaw, Charlton et al. 1988; Bailey and Gatrell 1995). The process is a computationally intensive approach to the automated identification of clusters.

GAM is a cluster location tool used to analyze spatial point distributions. Its purpose is to find evidence of localized geographical clustering. It incorporates a technique that compares the intensity of events within circles of varying radius; the circles are centered on a fine grid imposed over the area of interest. The levels of intensity are compared to a constant that represents the expected intensity. The circles identify areas of significant differences from the constant. The technique involves a Poisson model for the statistical distribution. Once the condition of statistical significance and the intensity rate is met the circle is drawn on the map over the area. This process is continued across the study area (Openshaw, Charlton et al. 1988).

This process only looks for the clusters and does not seek to explain the underlying cause of the cluster. Also by looking at the entire study area it does not require the

specification of a spatial scale at which we expect clusters to occur. Therefore this process makes few assumptions about the dataset and is considered an exploratory method (Bailey and Gatrell 1995).

2.4 Data Preparation/ Methods for Statistical modeling

After researching the appropriate spatial analytic tests, ClusterSeer™, a commercial software package, was selected because it provided a comprehensive suite of models for spatial data. ClusterSeer™ was specifically designed to analyze disease clusters in both spatial and temporal data. This program also was able to run all desired statistics except the Geographic Analysis Machine. The following sections are the methods for preparing the data for use with ClusterSeer™, and the processes involved in performing the models.

2.4.1 Ripley's K

Data preparation for running the Ripley's K model began with the entire dead bird point file. Selections were made for each month from the dataset using the Caliper GIS software TransCAD (Caliper 2006). These files were exported as shape files and used in ClusterSeer™ for processing. The ClusterSeer™ interface required a point file for this model with spatial location information and a unique ID field. The data were imported with a geographic projection (latitude and longitude), and kilometer was selected as the unit for distance. Selecting the proper projection information is important; Ripley's K reports an interpoint distance that allows for an interpretation of approximate cluster scale. If the wrong projection information were selected the units of the results may not be correct. Interpoint distance refers to the measurement between data points. Next,

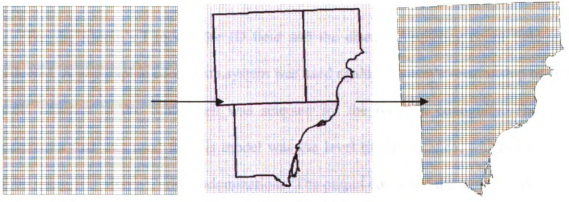
parameters for the distance step limits were set for the model to run. The distance step limit was set to 10 steps in each model. The distance used for each step was set as default; the software estimated an approximate minimum interpoint distance. It is required for determining the autocorrelation between points. Finally, the number of Monte Carlo Randomizations was set to 100. This model produced the output as the maximum deviation from identity (one-to-one line). It also produced a graph of the simulation envelope and *Lhat* line (Fig. 10-16).

2.4.2 Moran's I and Local Moran (LISA)

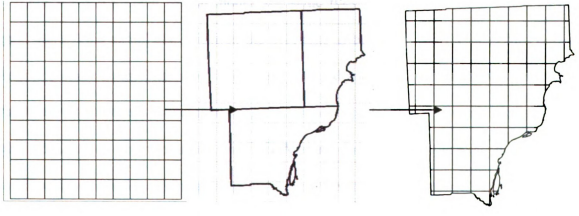
For the Global and Local Moran's I models the same 100 by 100 and 10 by 10 grid cell files were used. These models required a uniquely identified polygon file. The only attribute that was required for the models was a count of the disease records for each polygon; dead birds were used as the disease frequency. The polygon file had to be clean, with no double lines or overlapping segments, the background files that were available did not meet these requirements. For the purposes of being uniform and unbiased, the data were converted to a grid. To create the attributed grid the Caliper GIS product, TransCAD was used.

The data files that were used included the dead bird records, as points, clipped to the study area, and the geographic outline of the study area. TransCAD allows the import of ESRI shapefiles into the Caliper native format of the Standard Geographic File. Both files were imported into this GIS. The function used to make the grid was Create Vector Grid. This function allows for the creation of area grids with a specified number of cells or cells at a set distance. This allowed the entire map window to be covered with a set number of cells without specifying a size for the individual cells (Figs. 7 and 8). The

grids cover the entire map window going beyond the borders of the study area. This was corrected by using the Clip by Area function, leaving a grid the exact dimensions of the study area (Figs. 7 and 8). Two grid size resolutions were used for these models, 100 by 100 and 10 by 10. The 100 by 100 grid cells measure approximately 900 meters square, and the 10 by 10 grid cells measure approximately 9000 meters square. After the clipping process the numbers of cells for each resolution were, 6,019 and 79 respectively (Figs. 7 and 8).



100 by 100 grid processing steps. **Fig. 7**



10 by 10 grid processing steps. **Fig. 8**

The next step involved with the data preparation was attributing the grid with counts of the dead birds within each cell. A new field was added to the grid database. The field was filled using an aggregate function that counted the specified contents of the given cell. This worked extremely well. There was an issue of points being counted in more than one polygon using this function with other background files. This multiple counting did not occur with the grid base files. Summing the count column, then comparing to the total dead bird count, verified this. This was done for all of the bird records and for each of the time frames mentioned in (Table 2). The grids were then exported as shapefiles for use with ClusterSeer™.

For the Moran's I model the ID field and the disease frequency count field were selected. A geographic coordinate system was used again. A queen's relationship of eight cells adjacent to the center cell was selected for the polygon contiguity. The other parameter that was used for this model was the level of significance, which was set at 0.05 using 100 Monte Carlo randomizations. The output contained the numeric Moran's I results, a graph of the Monte Carlo distribution, and a plot of the p-values against the Monte Carlo runs. No map was produced from this model.

For Anselin's Local Moran model, the same files were used at the 100 and 10 cell grid sizes. This model also required the polygons to be clean, and the dead bird data in numeric count form. The same fields were selected and the significance was set at 0.05 with 100 Monte Carlo randomizations. The queen's relationship was used. Products from the LISA model were the same as with the Moran's I model, with the exception of a map attributed with the local Moran statistic. There were some problems with exporting some

of the shapefile maps from ClusterSeer™. The program had a tendency to freeze and shut down. However exporting the session logs presented no problems for any model.

2.4.3 Oden's I(pop) and Kulldorff's Scan Method

Data preparation for these models was very similar to the previous models. The additional information required for these statistics was the population information. Population was calculated for each cell in the grid. Some of the cells did not contain a Census block group point. ClusterSeer™ required a number greater than zero for all cells, to perform the models. For these blank cells a value of one was entered as the population count. Only four cells required this adjustment, and all four cells had a zero bird count. Therefore, the results were modestly affected. However, due to the non-zero requirement these models were not run using the 100 by 100 grid.

The last global model was Oden's I(pop). Oden's I(pop) required the unique ID field, a count field for the records in each grid cell, and a record for the aggregate population for each cell. With these fields selected, the projection was set. For this model the queen's relationship was selected. The significance threshold was set to an alpha of 0.05 and 100 Monte Carlo simulations were used. This model produced results for I(pop) and the significance level.

The final model was the Kulldorff Spatial Scan method. The Kulldorff Spatial Scan is a local cluster indicator, and has the option to look for clusters in space or space and time. The same three fields and projection information was entered. Also the queen's relationship was used for this model. The only other parameter for Kulldorff was the number of Monte Carlo randomizations, which was set at 100. Kulldorff also calculated

an average disease frequency based on the dead bird counts and population data. The advantage of this model is that it produces three classes of likely clusters and also provides a map showing the locations of each class. It also calculates average disease frequency, the log likelihood ratio and the upper-tail P-value for each class.

2.4.4 Geographic Analysis Machine

The Geographic Analysis Machine (GAM) was the only model that could not be completed using the ClusterSeer™ software program. The GAM model was programmed using the 'R' Statistical package. This programming language and analysis environment is used for statistical computation and graphics applications (Team 2006). The 'R' language is very similar to the 'S' language that was developed at Bell Laboratories (Chambers et al 2006).

The environment for executing the program requires libraries or packages for processing the various functions within the code. There were four libraries required for completion. The first was the splancs 2.01-17 package (Rowlingson and Diggle 2006). This library provided the spatial and space-time point pattern analysis functions. Next, the gstat 0.9-31 package was used for geostatistical modeling, prediction and simulation (Pebesma 2006). The third library was fields 2.3 (Nychka 2005), required for calculating the distance functions within the program. The last package, maptools 0.5-12 (Lewin-Koh and Bivand 2006), was required for reading in the boundary file, which was in "shapefile" format.

The study area boundary and population centroids were projected to UTM Zone 17 using the WGS 1984 datum. The dead bird point files that were used previously were re-

projected using TransCAD and exported as comma delimited text format. Also, the Census population centroids were converted to comma delimited text format. The boundary file was used as a convex hull for the data and in the final maps for this model.

After data import, the program first calculated a background rate for use in the model. The background rate was calculated by dividing the length of the dead bird point column and the sum of the population centroids. This value changed for every model because each time period had a different number of birds. Also required for the model was a grid of points overlaying the study area; the grid acts as the locations at which regional counts will be aggregated. Therefore, a finer resolution of grid points would make for a smoother final map. The finer resolution would also exponentially increase computation time. For all time periods a grid spacing of 1609.344 meters (1 mile) was used. Then the boundary file was used to remove the grid points outside the study area. GAM required a function to calculate the distance between the grid points and the bird points and the population centroids. The function used the Euclidian distances between the points for calculations. A four-column array was created to hold the results from the GAM calculations. The first column contained the number of birds within each calculated distance of each grid point. The second column contained the sum of the population centroids within the specified distance of each grid point, which only needed to be large enough to contain a bird and a population centroid. A distance of 10,000 meters was used. The third column was filled with the ratio of birds to people, and the fourth column contained the expected bird count, based on the population and background rate. With the array completed the confidence interval needed to be set. This method is known for requiring extremely high confidence intervals to see meaningful results. All models were

compared using the Poisson distribution starting at an alpha of .01 and increasing from there. The confidence interval was increased until there were only a few clusters present. The model produced a map of the study area showing areas of potential clustering for each time period.

3 Results & Analyses

3.1 Results from Global Methods

3.1.1 Ripley's K

The Ripley's K function models the nearest neighbor limited to the smallest possible scales, as though there is no first order effect within the dataset. In order to more easily interpret the results from the Ripley's K model as a straight line, the *Khat* statistic is converted to *Lhat*. This allowed for comparisons to a straight line. As the *Lhat* differs from the straight line (randomness) the point pattern can be interpreted as having spatial structure.

Examination of the graphs produced by the model showed a noticeable pattern associated with the amount of dead bird locations from each model. As the number of points increased the graph showed increasing spatial structure. This could be seen when comparing the graph for all results (Fig. 10) and the graphs for July and August (Fig. 13 and 14). These graphs showed a smooth arc for the data points far above the simulation envelope, denoting structure within the point pattern. The model for April showed no results due to insufficient points (two records). When reviewing the results for May (Fig. 11) the model showed complete spatial randomness. As the spatial relationships between the points change throughout the transmission season greater structure is present. This can be seen in the June data (Fig. 12); as the points began to show signs of positive structure *Lhat* is above the simulation envelope. The positive structure continued to intensify through the peak of the transmission season in August (Fig. 14). The data began

to lose cluster associations at the end of the season. The graph for September (Fig. 15) showed the model just outside of the prediction envelope. At the end of the season in October (Fig. 16) the model falls completely within the prediction envelope showing a return to randomness in the data. For this model and all subsequent models randomness describes a spatial pattern in which all records have the same probability of occurrence at any location. Also, a cluster is defined as a statistically significant subset within a population.

For a better understanding of the graphs a legend has been included (Fig. 9). The legend shows the identity function as the red line; this represents the one-to-one line signifying complete spatial randomness for the model. The gray lines represent the simulations run by the model; the combined extents of the simulations are shown as the envelope in blue; the average for the simulations is depicted in green. Together these lines show the extents of the model for randomness. The black line shows the fit of the data to the model. Interpoint distance is measured in kilometers.

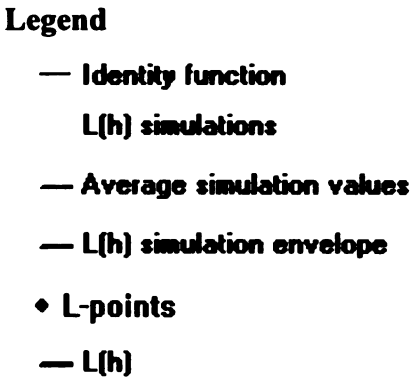


Fig. 9

All Records

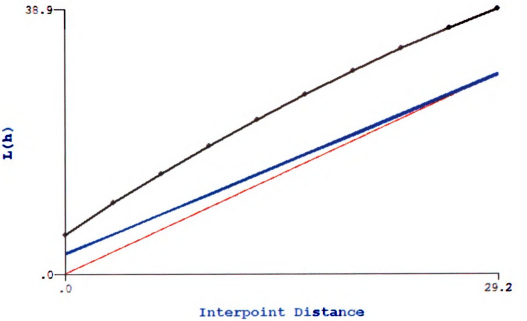


Fig. 10

May

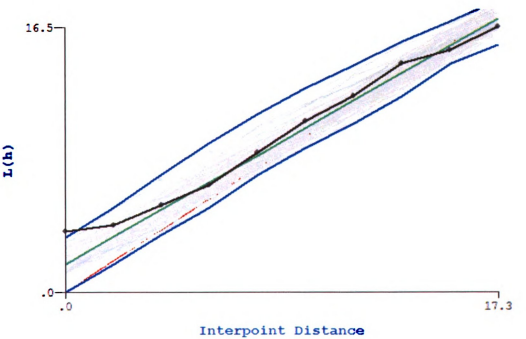


Fig. 11

June

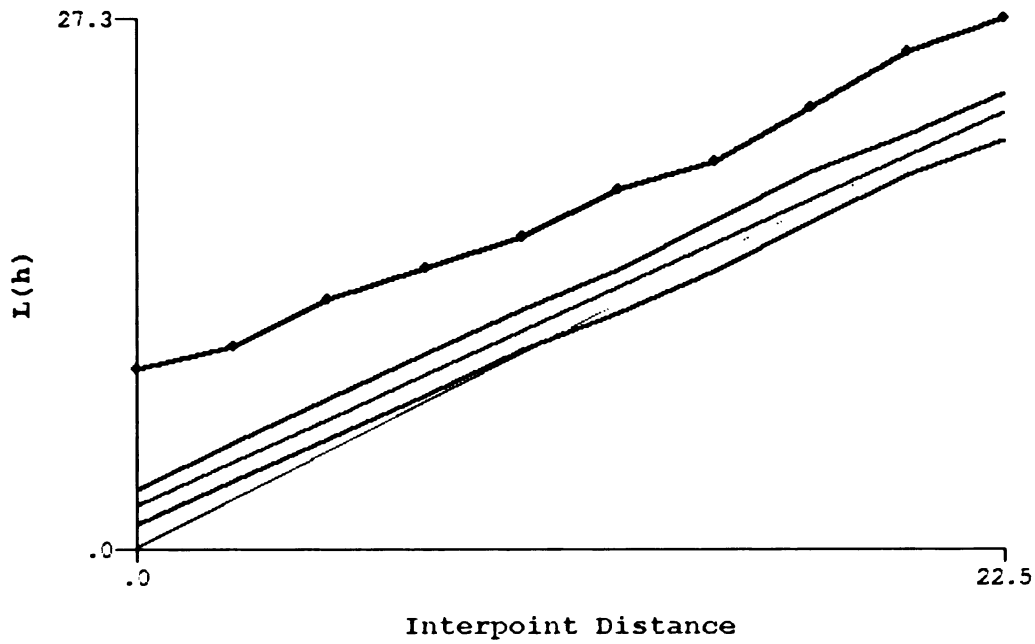


Fig. 12

July

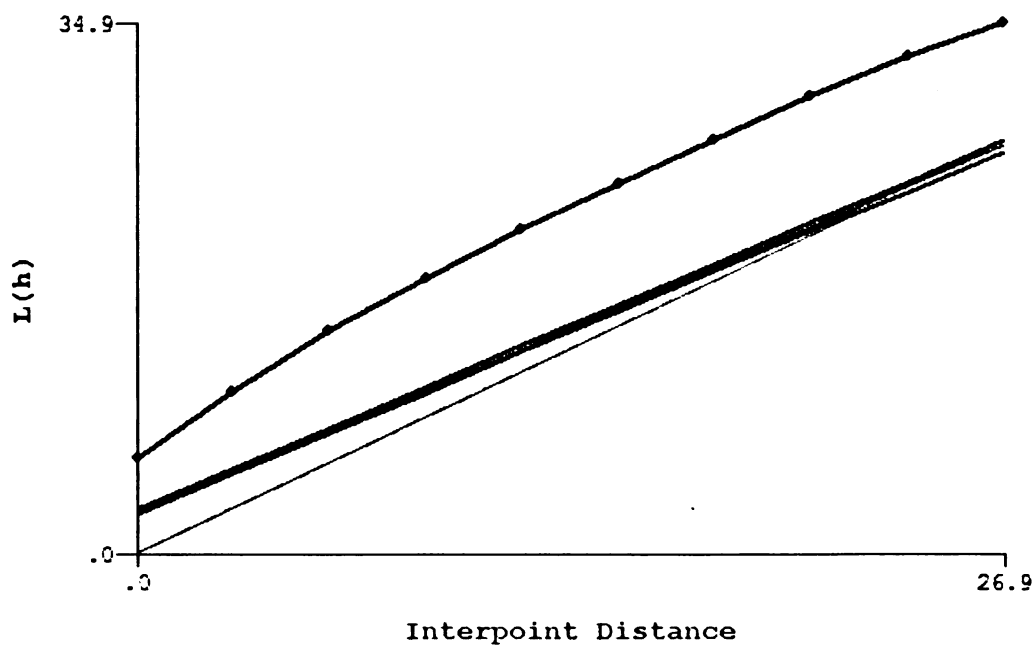


Fig. 13

August

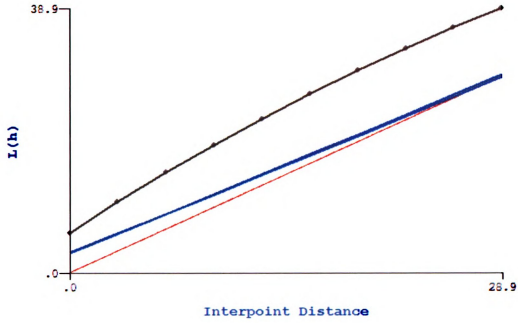


Fig. 14

September

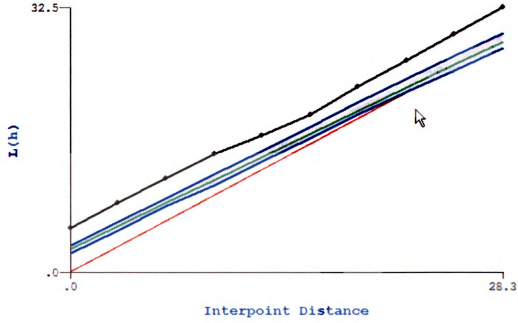


Fig. 15

October

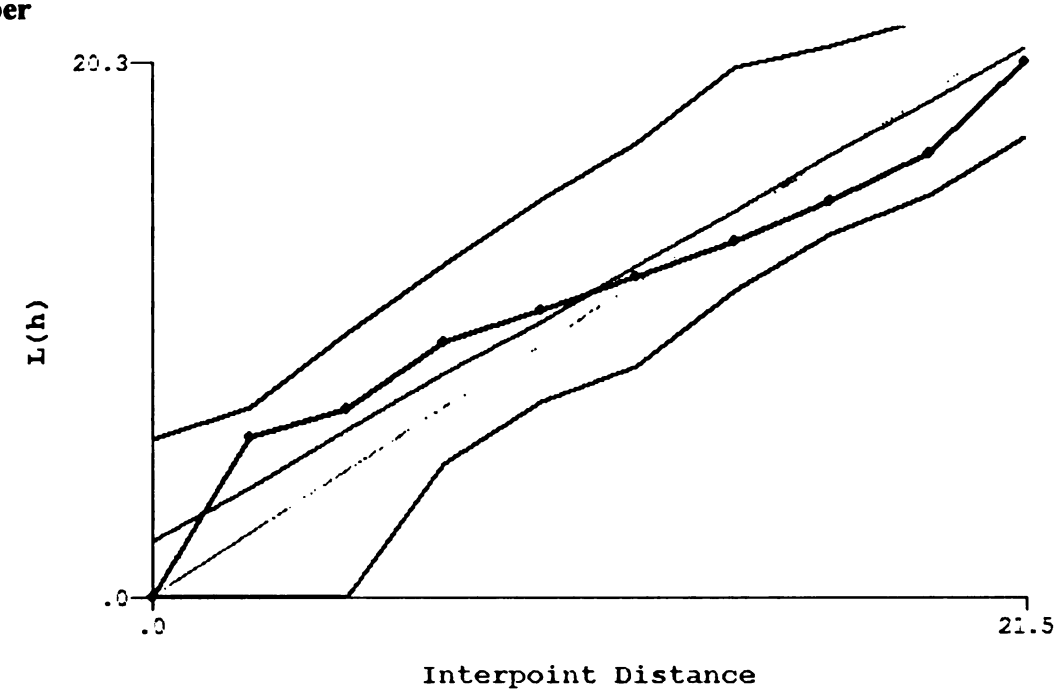


Fig. 16

3.1.2 Moran's I

Moran's I is a weighted correlation coefficient used to detect departures in spatial randomness in global level data. The statistic indicates spatial patterns such as clusters or geographic spatial trends across the area of interest. The values of Moran's I range from −1 to 1. Positive values represent positive spatial autocorrelation meaning that the neighboring areas contain similar rates. This relationship indicates spatial structuring within the point distribution. When Moran's I is negative the model indicates that the nearby areas are dissimilar, usually indicating an increase to a regular pattern.

The 100 by 100 cell grid was modeled first. For all dead bird records the model indicated a Moran's I value of 0.32 showing positive autocorrelation in the point distribution. From all the other time periods modeled, August indicated the greatest Moran's I value of 0.29. July was the only other month to show any significant positive

autocorrelation with a Moran's I value of 0.14. Values from outside the peak season were all very close to zero for other time periods. The significance values for these models only met the alpha value for the all records, July and August. Full results for this model can be found in Appendix 1.

When the data were aggregated to the coarser 10 by 10 grid there were many more models that showed spatial autocorrelation. The model for all records increased to 0.63 showing positive spatial autocorrelation. Early in the season the Moran's I value showed a small peak in May rising to 0.27. There is a brief drop-off during June, with a sharp increase toward the later half of the month. The Moran's I value increases through July showing increased autocorrelation with the season peak in the first part of August with a value of 0.64. The model for August showed the greatest positive autocorrelation Full results for this model can be found in Appendix 1.

3.1.3 Oden's Ipop

Oden's Ipop is an adaptation of Moran's I that has been adjusted for population. This model reports three different significance values using z-scores and variance, Monte Carlo randomization, and using multinomial randomization. A multinomial distribution describes the outcomes of independent trials with two or more possible but mutually exclusive outcomes. This approach is used for redistributing cases of disease among spatially referenced sub-groups. The cases are distributed among the sub-groups at random where the probability of a case being placed in a particular group is proportional to the population at risk for that group. The first two methods for significance are designed for data which is normally distributed, because the data are not assumed to be

normally distributed the Monte Carlo significance values will be used for analysis. The statistic also calculates the I(pop) value which is used to see if the data present spatial structure. The I(pop) statistic will get large when there is clustering within a cell or between cells. The null hypothesis for this model is that there is no clustering in the data, this hypothesis is excepted as the I(pop) statistic approaches zero. The range for this statistic is dependent on the size of the population. For this model population is large, This will cause the range of I(pop) to be very small. This model showed significance in months that contained large bird counts (all records, and July through October) (Table 3), and with I(pop) values close to zero for all models, structure is said to be present if the model was found to be significant. The full results for this test can be found in Appendix 2.

	Ipop	Ipop'	E[I]	Test Statistic	Upper Tail
All Records	8.80E-05	0.142425	-2.47E-07	0.142425	0.0099
April-May	6.56E-07	0.077993	-2.47E-07	0.077993	0.18812
June	7.40E-06	0.507311	-2.47E-07	0.507311	0.0099
June 1-15	8.38E-06	0.997917	-2.47E-07	0.997917	0.44554
June 16-30	-2.98E-07	-0.0482642	-2.47E-07	-0.0482642	0.44554
July	2.60E-05	0.265002	-2.47E-07	0.265002	0.0099
July 1-10	4.81E-07	0.0971934	-2.47E-07	0.0971934	0.16832
July 11-20	4.46E-06	0.166938	-2.47E-07	0.166938	0.0099
July 21-30	2.35E-06	0.354976	-2.47E-07	0.354976	0.0099
August	8.33E-05	0.184354	-2.47E-07	0.184354	0.0099
August 1-10	6.72E-05	0.310556	-2.47E-07	0.310556	0.0099
August 11-20	2.58E-05	0.226085	-2.47E-07	0.226085	0.0099
August 21-31	4.06E-05	0.334904	-2.47E-07	0.334904	0.0099
September	2.01E-05	0.490066	-2.47E-07	0.490066	0.0099
October	8.69E-06	2.19463	-2.47E-07	2.19463	0.0099

Oden's I(pop) Monte Carlo assumption results. An alpha of 0.05 and 100 Monte Carlo Simulations were used for all time periods

Table 3.

3.2 Local Models

3.2.1 Anselin's Local Indicator of Spatial Autocorrelation

Anselin's Local Indicator of Spatial Autocorrelation is a Moran's test identifying local clusters in group-level data. Consistent patterns in the data started to develop after the first couple months of the year. Clustering was focused near the center of the study area (Birmingham/ Royal Oak areas) and along the coast (Grosse Pointes areas).

The All Records map (Fig. 17) showed this in a dark green band across the middle of the study area. Moving away from these areas the intensity of the clustering tends to diffuse. The records for April had no valid results; therefore they were included with the records from May for modeling (Fig. 18). This model showed signs of clustering in areas with greater human population. The transmission season began to intensify during June (Fig. 19), and there was a shift in the areas of clustering. The focus changed to the Livonia/ Westland area due to a large number of records in the last two weeks of June (Fig. 20). This cluster is offset by the randomness shown in the first half of June (Fig. 19). The July map (Fig. 21) demonstrated a cluster pattern that surrounded the city of Detroit. The densely populated suburban cities around Detroit showed mild clustering with moderate intensity in the cities of Royal Oak and Warren. This pattern varies little during the month of July (Fig. 22). The August maps showed the peak of the season, as well as, the majority of the records. Throughout the month of August the cluster intensity shifts from the immediate areas around the city (Fig. 26), to areas farther away. The map for the first part of August (Fig. 27) showed the most intense clustering between the Redford/Southfield areas, with moderate clustering in Birmingham. For the next map for August (Fig. 28) the most intense areas migrate north to Birmingham and Southfield and

east to the Grosse Pointe communities. The final map for August (Fig. 29) showed this trend continuing as the clusters present over the Farmington and Bloomfield Hills areas. This map also showed the first instance of clustering advancing as far north as Rochester/ Rochester Hills areas, and as far west as Novi. Although the intensity of the clusters slightly drops off, the clusters are more widespread and occur in more areas. September (Fig. 30) showed very mild clustering throughout the middle of the study area. The map for October (Fig. 31) presented some clustering in the east side of Oakland County.

All Records

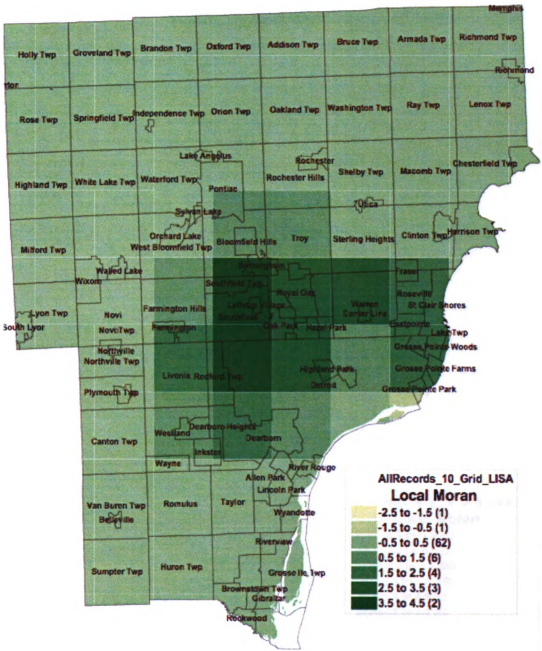


Fig. 17

April and May

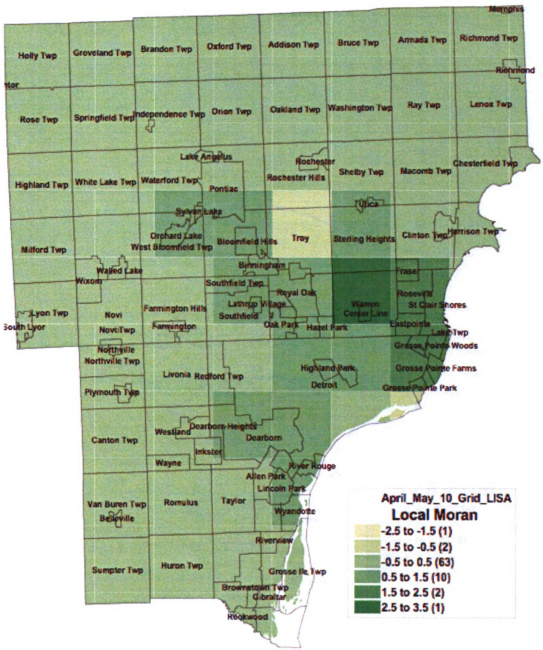


Fig. 18

June

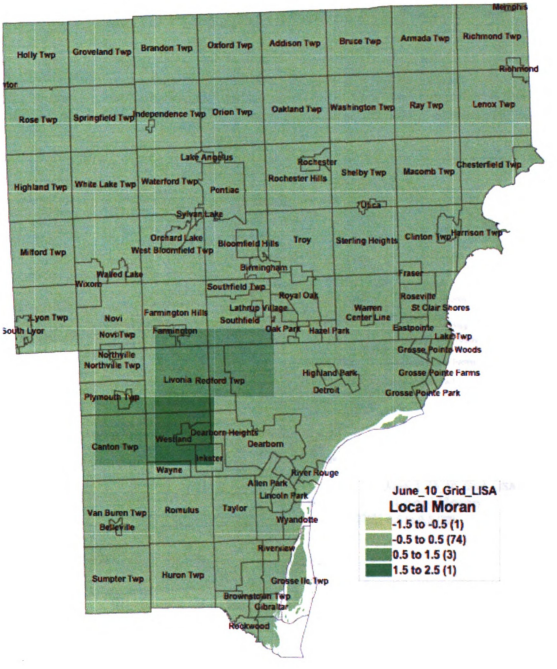

Holly Twp
Groveland Twp
Brandon Twp
Oxford Twp
Addison Twp
Bruce Twp
Armada Twp
Richmond Twp
Richmond
Rose Twp
Springfield Twp
Independence Twp
Orion Twp
Oakland Twp
Washington Twp
Ray Twp
Lenox Twp
Lake Angelus
Rochester
Rochester Hills
Shelby Twp
Macomb Twp
Chesterfield Twp
Highland Twp
White Lake Twp
Waterford Twp
Pontiac
Utica
Sylvan Lake
Orchard Lake
West Bloomfield Twp
Bloomfield Hills
Troy
Sterling Heights
Clinton Twp
Harrison Twp
Milford Twp
Walled Lake
Wixom
Birmingham
Fraser
Southfield Twp
Royal Oak
Roseville
St Clair Shores
Lathrup Village
Warren
Center Line
Lyon Twp
South Lyon
Novi
Farmington Hills
Southfield
Oak Park
Hazel Park
Eastpointe
Novi Twp
Farmington
Lake Twp
Grosse Pointe Woods
Northville
Northville Twp
Livonia
Redford Twp
Highland Park
Grosse Pointe Farms
Detroit
Plymouth Twp
Grosse Pointe Park
Dearborn Heights
Canton Twp
Westland
Dearborn
Inkster
Wayne
River Rouge
Allen Park
Lincoln Park
Van Buren Twp
Belleville
Romulus
Taylor
Wyandotte
Riverview
Sumpter Twp
Huron Twp
Grosse Ile Twp
Brownstown Twp
Gibraltar
Rockwood
June_10_Grid_LISA
Local Moran
-1.5 to -0.5 (1)
-0.5 to 0.5 (74)
0.5 to 1.5 (3)
1.5 to 2.5 (1)

Fig. 19

June 1-15

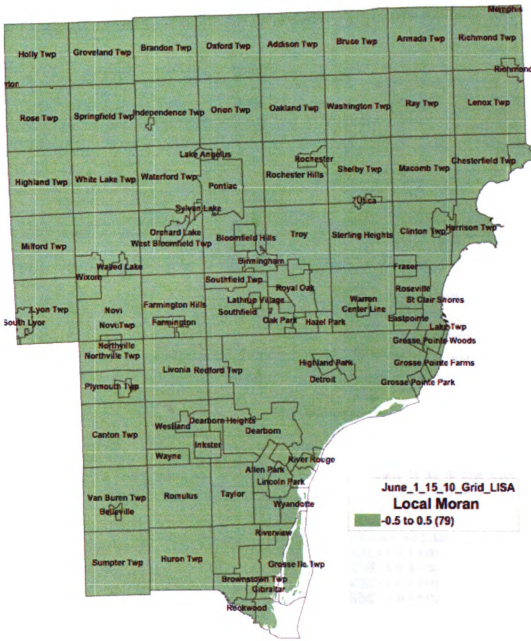


Fig. 20

June 16-30

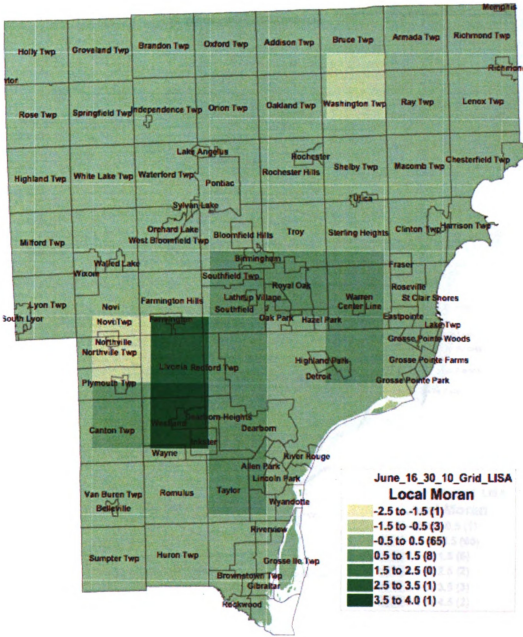


Fig. 21

July

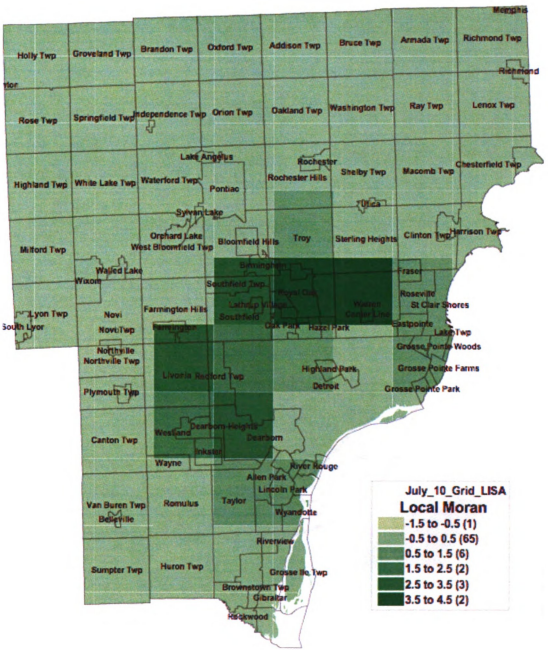

Holly Twp
Groveland Twp
Brandon Twp
Oxford Twp
Addison Twp
Bruce Twp
Armada Twp
Richmond Twp
Richmond
Rose Twp
Springfield Twp
Independence Twp
Orion Twp
Oakland Twp
Washington Twp
Ray Twp
Lenox Twp
Highland Twp
White Lake Twp
Waterford Twp
Lake Angelus
Pontiac
Rochester
Rochester Hills
Shelby Twp
Macomb Twp
Chesterfield Twp
Sylvan Lake
Milford Twp
Orchard Lake
West Bloomfield Twp
Bloomfield Hills
Troy
Sterling Heights
Clinton Twp
Harrison Twp
Walled Lake
Wixom
Southfield Twp
Royal Oak
Roseville
St Clair Shores
Lyon Twp
South Lyon
Novi
Novi Twp
Farmington Hills
Southfield
Oak Park
Hazel Park
Eastpointe
Lenox Twp
Northville
Northville Twp
Grosse Pointe Woods
Livonia
Redford Twp
Highland Park
Detroit
Grosse Pointe Farms
Grosse Pointe Park
Plymouth Twp
Canton Twp
Westland
Dearborn Heights
Dearborn
Inkster
Wayne
River Rouge
Allen Park
Lincoln Park
Van Buren Twp
Belleville
Romulus
Taylor
Riverview
Sumpter Twp
Huron Twp
Grosse Ile Twp
Brownstown Twp
Gibraltar
Rockwood
July_10_Grid_LISA
Local Moran
-1.5 to -0.5 (1)
-0.5 to 0.5 (65)
0.5 to 1.5 (6)
1.5 to 2.5 (2)
2.5 to 3.5 (3)
3.5 to 4.5 (2)

Fig. 22

July 1-10

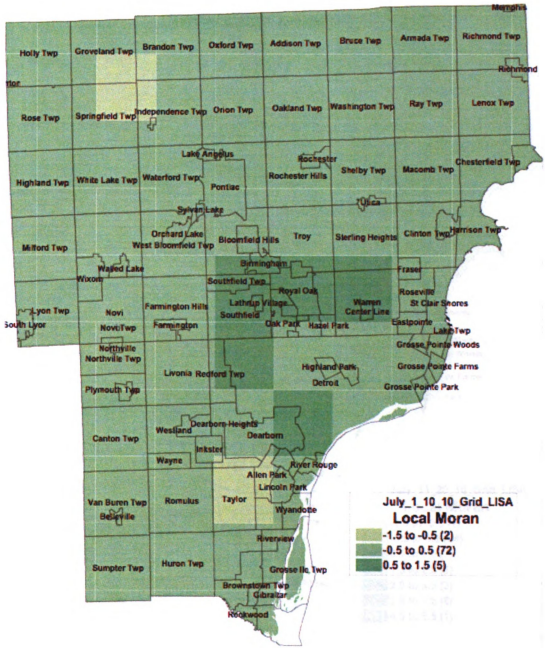


Fig. 23

July 11-20

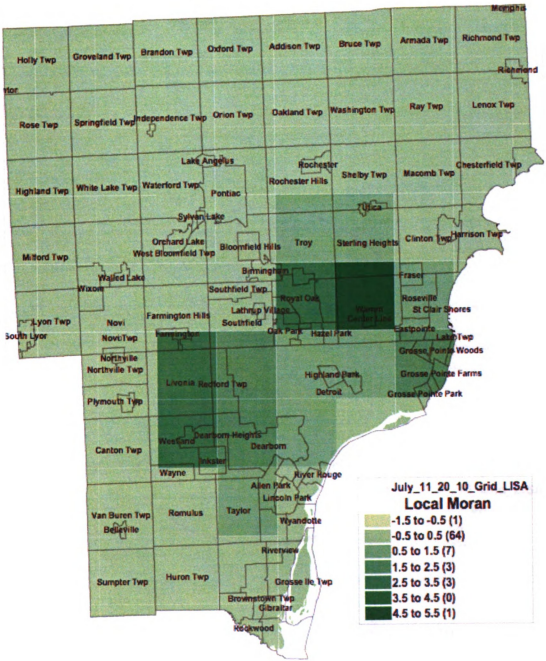


Fig. 24

July 21-31

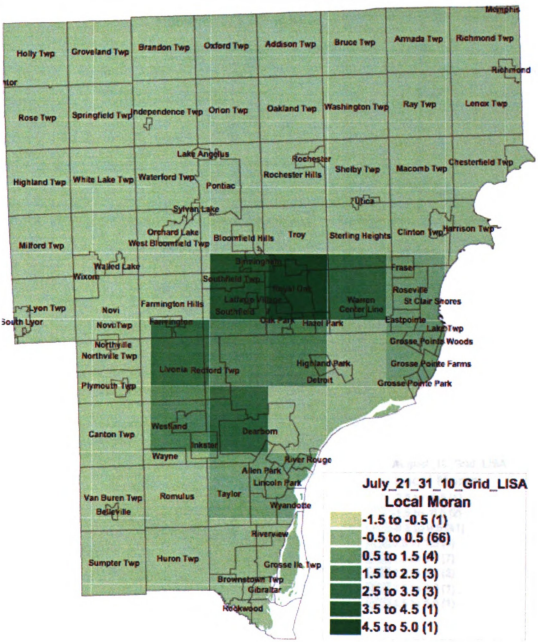


Fig. 25

August

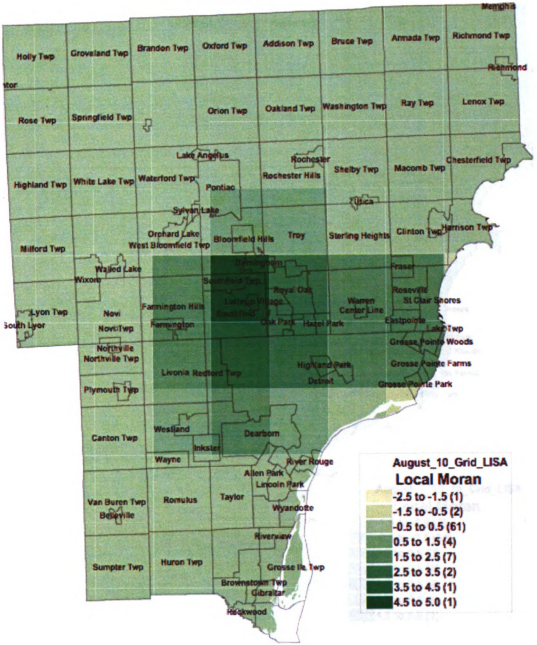


Fig. 26

August 1-10

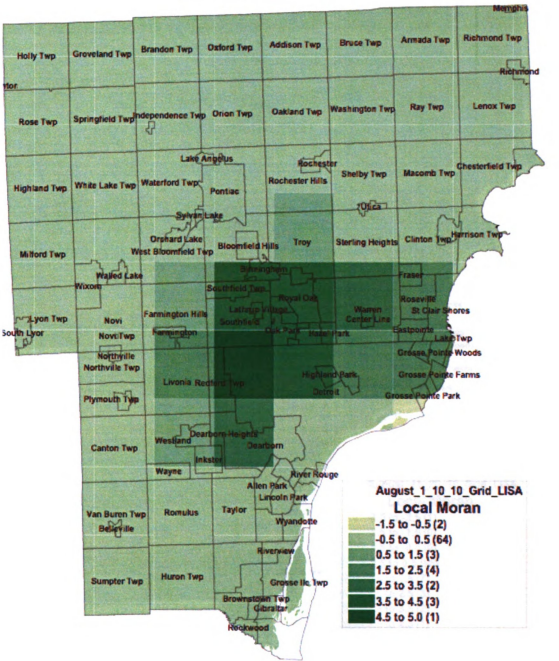


Fig. 27

August 11-20

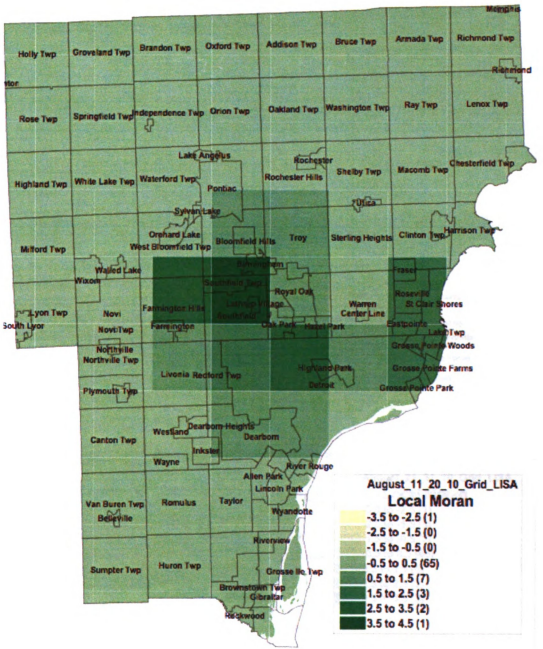


Fig. 28

August 21-31

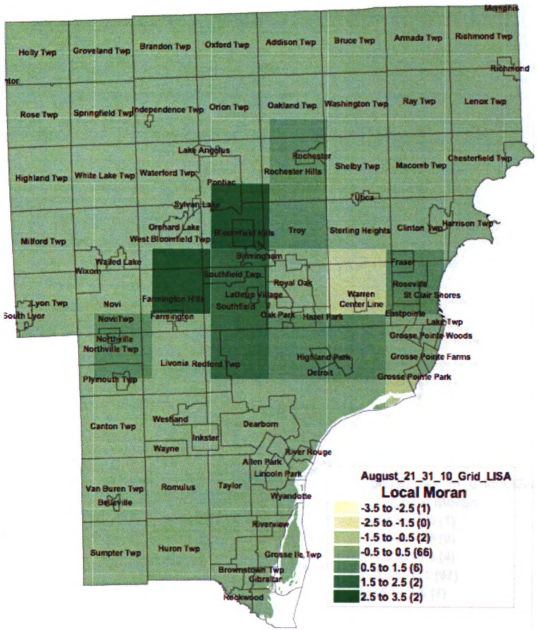


Fig. 29

September

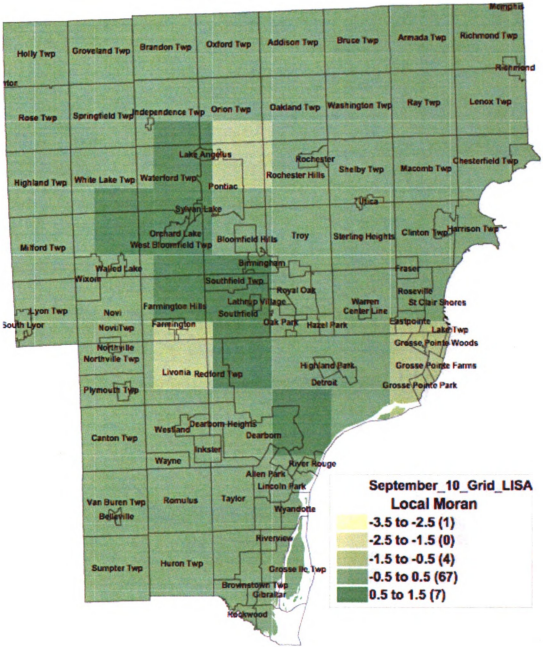


Fig. 30

October

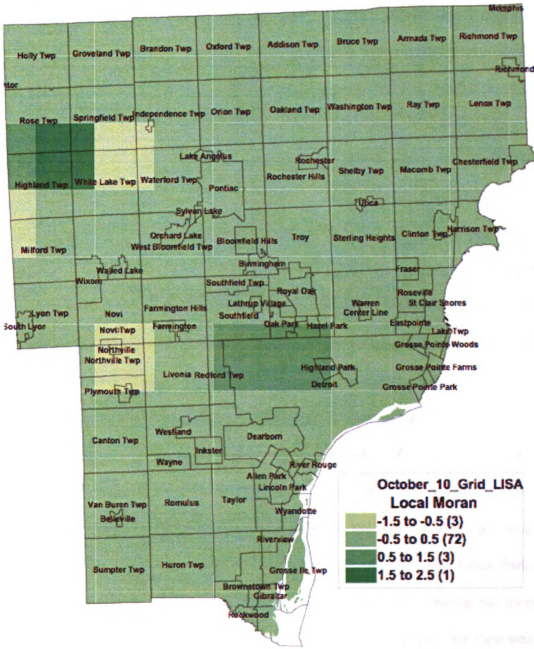


Fig. 31

3.2.2 Kulldorff

Kulldorff's scan method calculates three separate clusters ranked by likelihood for each time frame. "Likelihood" in this case refers to the chance of something actually being a cluster. The Kulldorff model identifies the three clusters based on the value of the test statistic, and whether the clusters overlap. The clusters are labeled as the first, second, and third most likely locations where a cluster may be present. By rule the second most likely cluster will not overlap the first, and the third will not overlap the second or the first. In the series of maps (Figs. 32-46), the first most likely cluster are represented in purple, the second most likely are shown in green, and the third most likely cluster are shown in yellow. The areas in blue showed no signs of clustering. When looking at the All Records map (Fig. 32) it showed that purple cluster for the dataset were in the center of the study area centered over the Birmingham/ Royal Oak/ Southfield areas. The green cluster was over the Grosse Pointe area, and the yellow cluster appeared just north of the purple cluster over Rochester. The rest of the plots showed results separated by month and the results widely varied across the study area. For the months of April and May the records were combined, due to insufficient data in April. The map for April and May (Fig. 33) showed the purple cluster to the north of the study area near Pontiac, with the green cluster located south near Lincoln Park and Allen Park. The yellow cluster was over Sterling Heights and Warren townships. During the month of June (Fig. 34) the clusters get much more focused. The purple cluster for June was near the Westland area, and the green cluster was just to the north over Livonia and Farmington Hills Township. The yellow cluster was a cell containing a single point separated from the rest of the points in the time period. Disregarding the one outlying

record, the map for the first half of June (Fig. 35) showed a pattern similar to the results for the whole month of June. The clusters focused around Westland, Farmington Hills, and Highland Park. The second half of June (Fig. 36) presented a large area on the west side of the study area as the purple cluster, with the single outlying point as the green cluster, and a small cluster around the Grosse Pointes, shown as yellow. For July (Fig. 37), the dead bird cases began to increase in density. The purple cluster for July was a large area covering much of the northeast part of the study area from Southfield to Shelby Township. The green cluster was located near the same area as the green cluster for the month of June (Livonia and Farmington Hills), and the yellow cluster covers Redford and Dearborn. The data records from the first part of July (Fig. 38) showed the purple cluster located over Independence Township in the northwest corner of the study area. The green cluster was over Taylor Township, and the yellow cluster was a large area reaching from St. Clair to Royal Oak and as far north as Shelby Township. For the middle of July (Fig. 37), the clusters continued to be located in similar areas as previously modeled time frames, showing clusters over Sterling Heights, Rochester, and Livonia. Toward the end of July (Fig. 40) the pattern began to reflect similarities to the All Records map (Fig. 32), showing a large cluster over the center of the study area with smaller clusters near St. Clair Shores. Throughout August the pattern of clustering more closely resembles large singular cluster shown in the All Records map. All maps for August (Fig. 41-44) maintain one strong, centralized cluster with a smaller, less significant cluster over the Grosse Pointe communities. During September (Fig. 45) the point pattern was less dense and caused a large dominating cluster over the northwest region of the study area. However, there are still signs of clustering over the Grosse Pointe communities. For the

October records (Fig. 46) there are signs of strong clustering over the northwest region of the study area, and along the coastline north of Harrison Township.

All Records

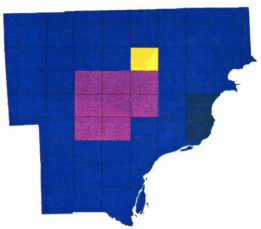


Fig. 32

April and May

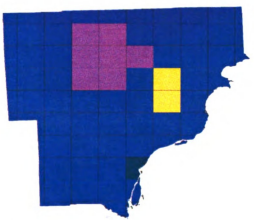


Fig. 33

June

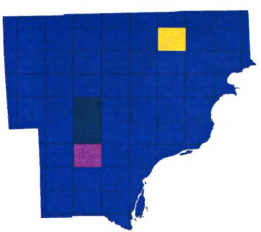


Fig. 34

June 1-15

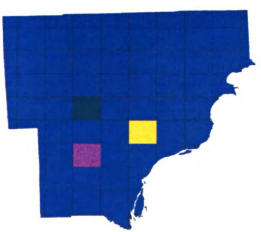


Fig. 35

June 16-30

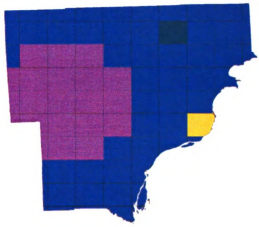


Fig. 36

July

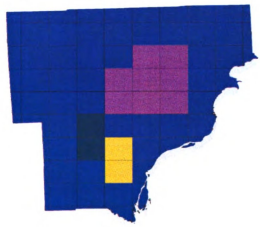


Fig. 37

July 1-10

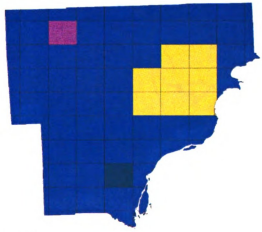


Fig. 38

July 11-20

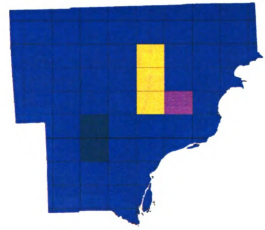


Fig. 39

July 21-31

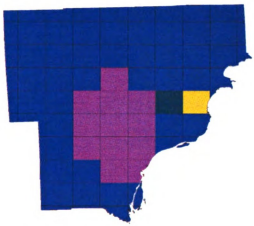


Fig. 40

August

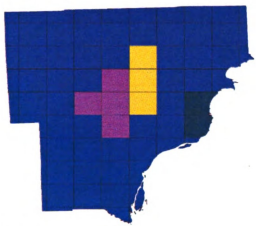


Fig. 41

August 1-10

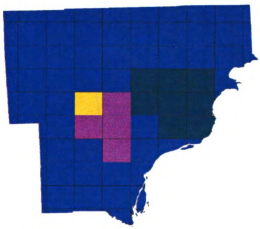


Fig. 42

August 11-20

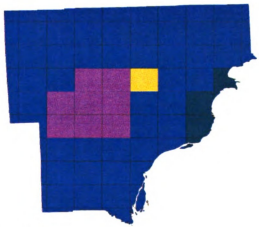


Fig. 43

August 21-31

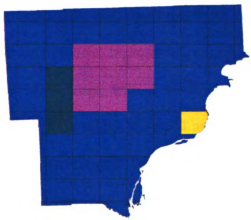


Fig. 44

September

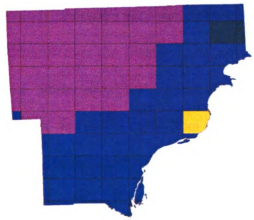


Fig. 45

October

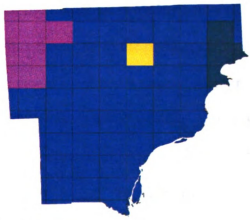


Fig. 46

3.2.3 GAM

The outputs from the GAM model required setting the confidence interval extremely high to begin to focus the areas of greatest cluster intensity. The models were all initially completed with a 99% confidence interval; however this caused large areas to be identified as significant. Therefore, higher confidence intervals were used to narrow the clusters to more focused areas. Due to the number and locational variability of the data

records in each time period and the underlying population in these locations, different confidence intervals were required to achieve the greatest focus of the cluster detection.

The All Records file (Figs. 46-49) for this model showed a very similar pattern to the previous models. It presented a large central cluster over the Birmingham/ Royal Oak/ Southfield areas, with a smaller and separate cluster over the Grosse Pointes. The decision to combine the April and May records for this model came from the results of the previous models. The combination of these records from early in the year showed two small clusters. The larger of the two clusters was centered over Pontiac Township and the smaller cluster was located over the Grosse Pointe communities (Fig. 50). The maps for June initially showed one large cluster over the southern part of the study area. However after increasing the confidence interval, most of the intensity was over the Westland area, with a thin band to the north reaching into Livonia (Figs. 51-53). For the month of July, the model identified a large cluster stretching from Birmingham to St. Clair Shores, and another large cluster reaching from Livonia to Westland (Figs. 54-63). The results for August show similar patterns to the All Records maps. Two distinct clusters for the month of August were identified. The larger more intense cluster centralized over the Birmingham area, and the second cluster covered the Grosse Pointes (Figs. 64-75). For September clustering was identified over the north half of the study area. Initially the clustering was widespread; however after increasing the confidence interval the clusters could be focused to two locations near Independence and Washington townships (Figs. 76-78). With few data points in October, the clusters moved even further into the rural areas of the study area. The main cluster was located over Highland Township (Fig 79-81); this is consistent with the results from the LISA and the Kulldorff models.

Potential GAM Clusters for All Records

99% Confidence Interval

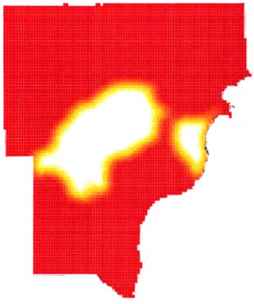


Fig. 47

99.999% Confidence Interval

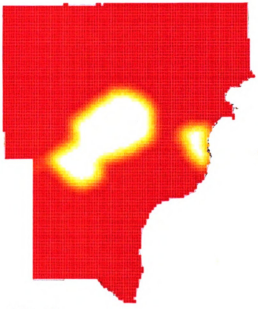


Fig. 48

99.99999999% Confidence Interval

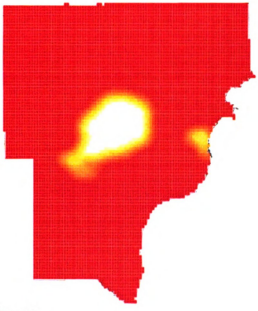


Fig. 49

Potential GAM Clusters for April and May

99.9% Confidence Interval

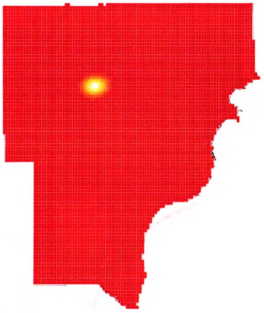


Fig. 50

Potential GAM Clusters for June

99% Confidence Interval

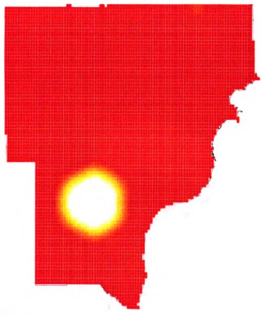


Fig. 51

99.999% Confidence Interval

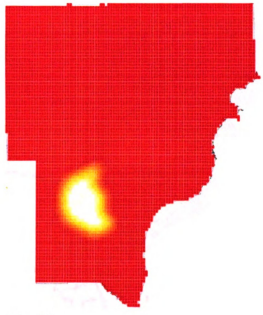


Fig. 52

99.99999% Confidence Interval

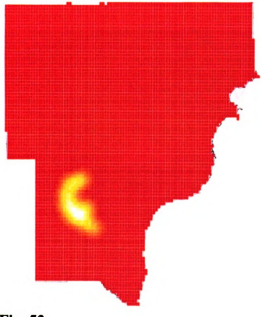


Fig. 53

Potential GAM Clusters for July

99% Confidence Interval

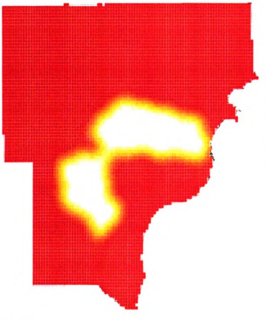


Fig. 54

99.999% Confidence Interval

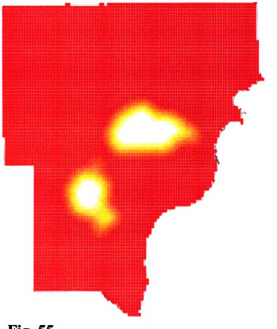


Fig. 55

99.99999999% Confidence Interval

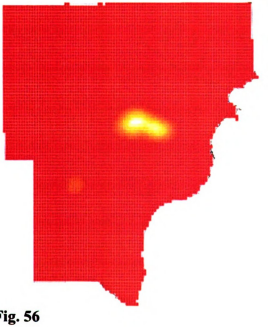


Fig. 56

Potential GAM Clusters for July 1-10

99% Confidence Interval

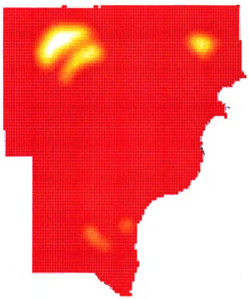


Fig. 57

99.9% Confidence Interval

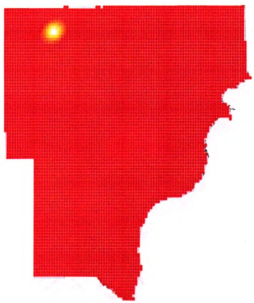


Fig. 58

Potential GAM Clusters for July 11-20

99% Confidence Interval

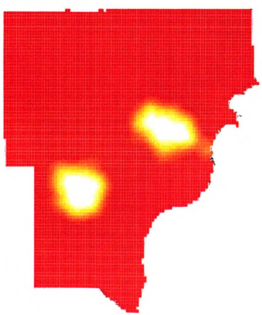


Fig. 59

99.999% Confidence Interval

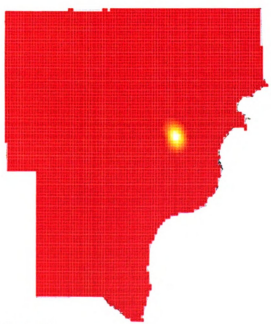


Fig. 60

Potential GAM Clusters for July 21-31

99% Confidence Interval

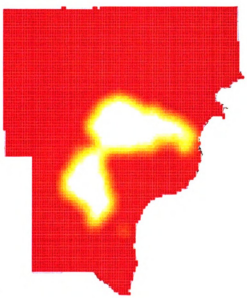


Fig. 61

99.999% Confidence Interval

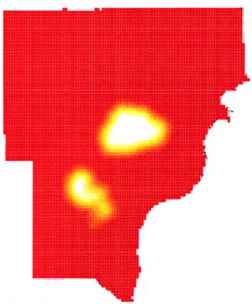


Fig. 62

99.99999999% Confidence Interval

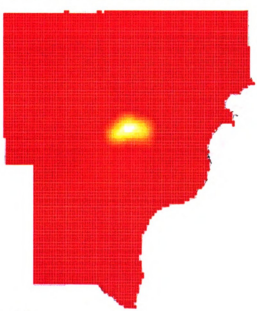


Fig. 63

Potential GAM Clusters for August

99% Confidence Interval

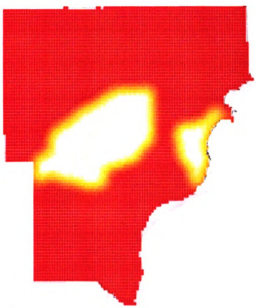


Fig. 64

99.999% Confidence Interval

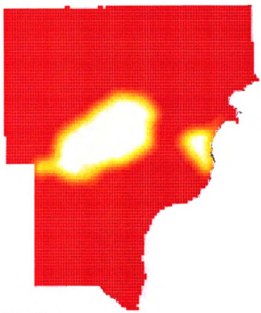


Fig. 65

99.99999999% Confidence Interval

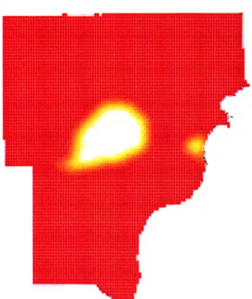


Fig. 66

Potential GAM Clusters for August 1-10

99% Confidence Interval

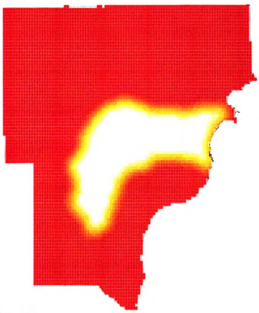


Fig. 67

99.999% Confidence Interval

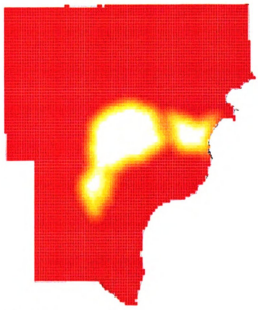


Fig. 68

99.99999999% Confidence Interval

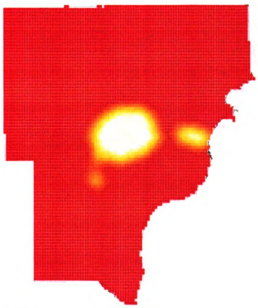


Fig. 69

Potential GAM Clusters for August 11-20

99% Confidence Interval

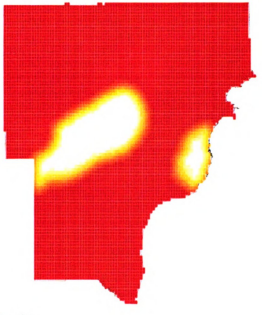


Fig. 70

99.999% Confidence Interval

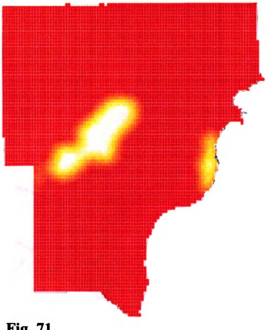


Fig. 71

99.99999999% Confidence Interval

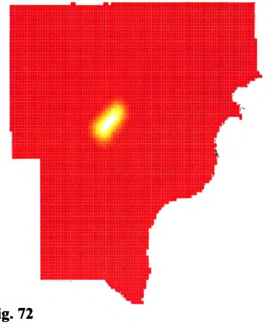


Fig. 72

Potential GAM Clusters for August 21-31

99% Confidence Interval

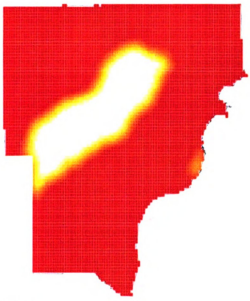


Fig. 73

99.999% Confidence Interval

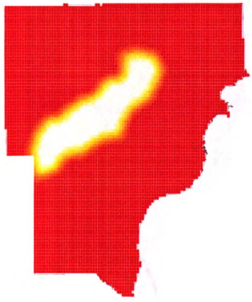


Fig. 74

99.99999999% Confidence Interval

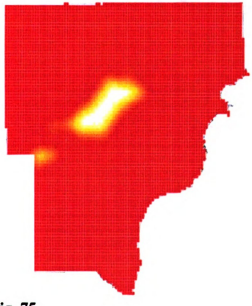


Fig. 75

Potential GAM Clusters for September

99% Confidence Interval

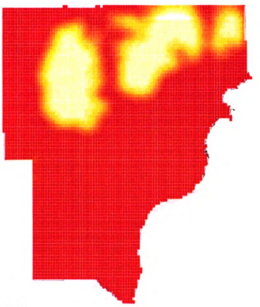


Fig. 76

99.999% Confidence Interval

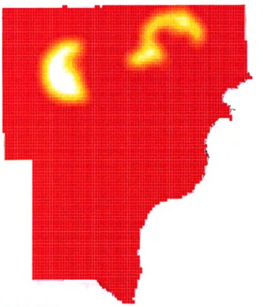


Fig. 77

99.99999% Confidence Interval

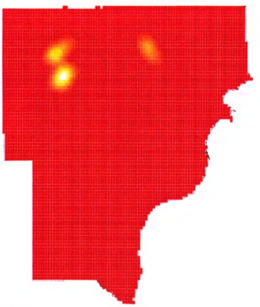


Fig. 78

Potential GAM Clusters for October

99% Confidence Interval

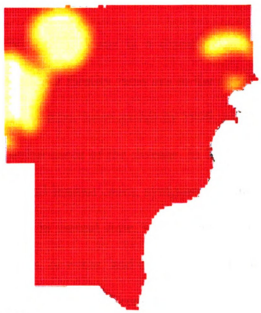


Fig. 79

99.9% Confidence Interval

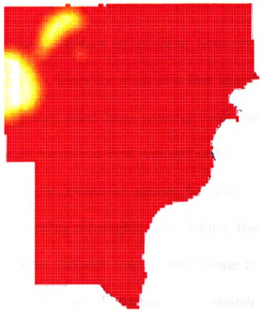


Fig. 80

99.99% Confidence Interval

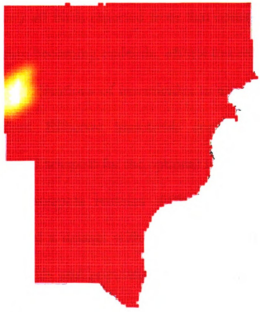


Fig. 81

4 Discussion and Conclusions

4.1 Overview

West Nile Virus is a disease that affects many species of animals; however the highest case fatality rates are in the avian population. Since the virus was first identified in the western hemisphere in 1999, all 48 contiguous states have reported both human and avian cases of the disease. In 2001, West Nile Virus was first reported in Michigan. In clinical human cases, symptoms exhibited are fever, weakness, nausea, headache, and changes in mental state, with advanced cases developing meningitis, encephalitis, flaccid paralysis, or coma, and occasionally death. The most common host for the disease is the American Crow (*Corvus branchyrhynchus)*. The Michigan Department of Community Health (MDCH) collected the spatial locations of over 8000 dead birds during 2002. Initially the birds were tested for WNV; however once the disease was identified in a county all birds reported from that county were assumed to be WNV positive. With the disease becoming endemic, efforts toward minimizing the human impact have increased, with applications of insecticides to eliminate the mosquito vector the most widely used method. Effective insecticide application requires the identification of spatial and temporal hotspots for the optional insecticide dispersal locations. The ability to accurately locate hotspots remains difficult because the data were voluntarily reported. This produced a dataset with a direct correlation between the numbers and locations of the dead birds and population density. This research used a subset of the data that were collected in 2002 in the Detroit Metropolitan area.

The dead bird data were collected through an information hotline which the public would call and leave their home address and the type of bird that was found. Due to the large number of dead birds reported, testing on each bird could not be completed, which leaves the possibility that some reported birds may have died of factors other than WNV. However with over 8000 reports statewide, and the mass deaths in the bird population, any errors in the data attributed to reported bird deaths of other causes are thought to be negligible. The spatial locations were identified for the dead bird records in the database using address matching via the Michigan Geographic Framework from the Michigan Center for Geographic Information (CGI). This data layer has a reported spatial accuracy of +/- 20 feet. The attribute used identifies a range of addresses found on each segment of street; for example if the segment range is 1000-2000, and the reported address is 1500, then the point will be placed at the mid-point on the segment. This decreases spatial accuracy in the points by not taking parcel size and density into account; for example rural points could possibly be off by a mile or more. Road naming conventions are also an issue. Errors in the spelling of road names either from the bird reports or within the road framework create the possibility of placing points on the wrong street, county, or not placing them at all. The process of attributing spatial data to a GIS model propagates any error already in the data with the error in the associated data layers. Ideally, spatial location data would be collected using a Global Positioning System to accurately place these locations using real world coordinates without projecting, transforming, or matching the locations before analysis.

In this study, better correlations were obtained when the data were aggregated to a larger cell size. However, one should be aware that this may be related to problems

related to the modifiable areal unit problem (MAUP;(Qi and Wu 1996)). When working with spatial data, if one does not use the same size of spatial unit in two different analyses, either the window size, or polygons of the same size, results may not be directly comparable. Because of the MAUP, generally as the window size increases, correlations between two variables are expected to improve. In this study, larger cell size produced stronger correlations as expected to the MAUP, and when the finer grid cell size was used weaker correlations were produced as was expected relative to the MAUP. It would be ideal to be able to determine the magnitude of the MAUP effect in this study. Although a number of researchers have suggested ways to analyze the MAUP effect, the process generally employs stochastic simulation to generate a number of different window sizes/aggregation schemes to evaluate the range of correlations (Openshaw 1984; Jelinski and Wu 1996), these researchers also note that the magnitude of MAUP is unique to the phenomenon being studied and can not be removed. Therefore, conclusions of any studies that employ areal sample units must be constrained to the spatial units employed (Plante, Lowell et al. 2004)

4.2 Discussion of Results

4.2.1 Global Tests

Three separate models were employed to examine global spatial trends. Global trends examine patterns over an entire study area. Each model was run with the ClusterSeer™ software program using the dead bird data. These models were designed to show if the data presented an overall trend. These methods do not identify the locations of clusters, but rather only whether spatial structure is present.

Ripley's K model reported results with a graph (Figs. 10-16), which showed an envelope of Monte Carlo simulations representing spatial randomness, and the relationship of the envelope to the empirically calculated statistic. The Ripley's K model determines if structure is present in the point pattern, and approximates an interpoint distance. When the line representing the model is above the simulation envelope structuring is present, and when the model is below the simulation envelope the result is interpreted as regularity. The results can be interpreted across a range of scales. For the purposes of this model, the data were first modeled in its entirety then divided by month. There was a direct relationship between the deviation from randomness and the number of bird points. As the number of birds in each month increased, the model showed greater amounts of structure. The results for the Ripley's K model showed spatial structure, however from the graphs one cannot interpret the pattern as being clustered. Because *Lhat* is above the simulation envelope across all scales the results are interpreted as a first order trend and invalidate the model. Ripley's K is a good indicator of spatial structure in point level data, but the model does not specify whether there are many clusters, or one large cluster.

Moran's I was modeled at two different grid sizes, 100 by 100 and 10 by 10. The data points were aggregated to the grid cells as numeric counts. The results from the 100 by 100 cell grid models were significantly lower than the 10 by 10 cell models. By using more grid cells the data were dispersed, and each cell contained few, but similar, dead bird counts. Thus Moran's I showed no signs of spatial structure. By increasing the grid cell size, from 100 by 100 to 10 by 10, larger groups of points could be modeled for spatial structure. The larger grid cell models showed the same patterns throughout the

transmission season with an overall increase in Moran's I value. The increased cell size identified positive spatial structure during July and August where the dead bird counts were the highest. Both the Moran's I and Ripley's K models work well for identifying possible spatial patterns that may contain clusters. However these models only take into account the relationship between the points and not the cause or spatial distribution of the patterns.

To attempt to identify a cluster pattern with the underlying population information, Oden's I(pop) model was used. Oden's I(pop) is designed to determine if the Moran's I patterns are still present when population is incorporated. Oden's I(pop) was of specific interest due to the population-based collection method. Months with large bird counts were found to be significant; however the I(pop) statistic was extremely low for all months. The low values of this statistic are attributed to the extremely large human population count used in this model. Because of the low values, the I(pop) statistic may be precisely interpretable. However assumptions can be made on the relationship between the values. Since the population stays constant through the models and the dead bird count changes, the increase in I(pop) values can be interpreted as structure.

All of the global models presented results showing the dead bird dataset had signs of spatial structure. Cluster locations cannot be identified with global models. Global models also do not describe the size or if there is more than one cluster present in the data. However, the modeling process proved helpful. Ripley's K was run first to determine any significant trending in the dataset. Then Moran's I identified that global structure was present. Finally, Oden's I(pop) was modeled to determine positive spatial structure would be found significant by accounting for the human population. With these

initial models confirming that there was possible clustering in the dataset, the focus turned to identifying the locations of the clusters. For this information, local cluster models were employed.

4.2.2 Local Models

Three separate models were used to examine local clustering. Anselin's Local Indicator of Spatial Autocorrelation and the Kulldorff's Spatial Scan models were run using the ClusterSeer™ software program. The geographic analysis machine was programmed using the 'R' statistical program. These models were designed to identify if any local clusters were present, and to locate areas with a high likelihood of having hotspots.

The local Moran's model examined local clustering in group-level data, and produced results that indicated clustering in very similar parts of the study area over time. From late June through September (Figs. 21-30), cluster patterns were most intense over Birmingham, Grosse Pointe and Dearborn Heights communities. The overall results (Fig. 17) for this model show the highest degree of clustering in the Birmingham/Southfield and Grosse Pointe bedroom communities. These areas are some of the most affluent in the study area. The communities with higher incomes and improved access to community information, as well as homeowners working in there own yards could attribute to greater reporting of dead birds from a more informed population. The advertising methods used to inform the public of this reporting program are unknown; however, they would also have a direct effect on the patterns seen in these data. People who were not informed of

the program would not have reported dead birds. This may be why there are no clusters seen within the Detroit City limits, and lower income communities.

Kulldorff's Spatial Scan modeled the dataset for adjacent regions (using the queen's pattern) with strong autocorrelation, to indicate clustering. Kulldorff uses an increasing circular window to identify areas of clustering. The window has an upper extent of 50% of the size of the study area, creating the possibility of the cluster potentially being nearly as large as the study area. Kulldorff locates and ranks the three most likely clusters. The clusters for this model appeared to be much larger than the Local Moran test; however they were located in very similar locations. The larger clusters hold very similar shapes over the underlying grid. The circular window that this model uses can be seen in areas with large bird counts; therefore irregular clusters would be hard to identify. This circular pattern shows late into the transmission season covering large, more rural, parts of the study area. Because of the low bird counts and low populations in these areas, the shift of these large clusters to the more rural areas makes identifying fine scale local hotspots difficult. It almost appears as though the model attempts to include all possible points in the three most likely clusters. The grid cells in the outlying areas contain such few points and have similar population counts, that the model measures similar rates of people to birds. It then identifies these large areas as clusters, even though they may be the only dead bird points for that time period. Overall this model showed clustering in similar locations to the local Moran's I model with larger clusters.

The Geographic Analysis Machine allowed for local analysis on individual records and the inclusion of the population data. This was the only model run on the dataset to include population data and examine the dead bird locations as individual events. Oden's

I(pop), LISA, and Kulldorff aggregated the population data to the grid cells. GAM was first parameterized using a 99% confidence interval. This identified large areas as significant. By further increasing the confidence interval to 99.99999999% (Fig. 49), smaller areas were then identified. This focused the model on those areas with the highest degree of clustering. Although the clusters were not small enough to pinpoint something as small as a neighborhood, it did highlight community-sized hotspots such as Grosse Pointe. This test demonstrated that hotspots could be identified to the community scale of accuracy. The hotspots reoccurred in the same areas found in both the local Moran's I and Kulldorff models, confirming these locations of having the greatest dead bird to population ratio in the study area. Although Kulldorff and GAM identified similar cluster locations, the size and shape of the clusters varied. Kulldorff seemed to have a tendency to attempt to include all data points in three clusters; the GAM model ignored these outliers. The Kulldorff and GAM models used similar methods to scan the dataset (circular window over grid intersections), the difference between the models occurred in how the two models reported results. Kulldorff examined adjacent regions for similarities; if the regions were similar then a cluster may be present. The three areas with the most similarities were ranked as clusters. The reported cluster's shape was an artifact of the grid used; using a different grid cell size, or different shape polygons may have provided very different results. The GAM compiled circular regions of various sizes that had higher than expected dead bird-to-person ratios, these results were then smoothed for representation on the map. There was no limit to the number or the size of clusters that GAM produced. However, the shape of the clusters tended to have "half-moon" sections

removed as the confidence interval increases (Fig. 51-53). This pattern was due to the circular window no longer finding the cluster significant.

The methods used in this research were designed to test the utility of datasets that are reported by private citizens. By allowing the public to report information, many data problems that already exist in well thought out data collection efforts are exacerbated. Errors in data collection and data entry are always present, however when the data is collected by phone call and web-based form, the quality control is highly speculative. The data collector not only has to get the information provided correct, they also have to trust that the source has credible information. The ability to monitor and maintain an accurate database becomes suspect, as well as the accuracy of the information and spatial locations provided. This combined with the dataset being biased to areas where people lived and were informed of the program created a difficult data problem. In an attempt to examine how these issues affected local clustering in spatial datasets two standardized tests were run from a commercially available software program and one model was programmed using a statistical package. Despite the challenges presented in this dataset, similar results were found using each of the models. All local models identified clusters in similar areas. However, these areas also support large human populations, and high numbers of reported dead birds. The local models run on the complete datasets showed nearly the same overall pattern. One large cluster was present in all three of the local models for the complete dataset, showing that overall there is a single dominating cluster for this dataset. Only when these data are separated into small units of time is local clustering visible. The similarity of the cluster patterns show that the models did identify

hotspots within the dataset and further testing is needed to determine the dynamics of these patterns.

4.3 Future Research

Further research on this topic could follow two different tracks. First from the clustering perspective, alternative models and methods to identify spatial and temporal clustering in the dead bird data would be useful, such as Besag & Newell's method or Grimson's method. The second would examine the collection methods, environmental, and biological aspects related to the spread of WNV.

Advancing the cluster analyses of this research would require additional years of data collection. One additional year of data would allow for analyses on the reoccurrence of the spatial patterns found in this study, and provide the ability to examine the temporal clustering of the dead birds. Another year of data would allow for comparisons between seasons to determine any relationships in the overall peak and smaller peaks within the transmission season. One year of additional data would be the minimum required to examine if these patterns are consistent. Many more years of data would allow for studies on the movement of the disease, examining if the clusters tend to shift or change in size. The combination of additional data would allow for the patterns identified in each season to be jointly analyzed by randomly removing 20% (or the Case Fatality Rate) of the data points and examining if the patterns still show consistent clustering. The development of a control dataset with separate clusters containing disease rates similar to that of the dead bird dataset and testing the model against other statistical methods would allow for evaluation of the models used in this research.

Because the results for each model could not be directly compared from the statistical results, it was difficult to make any overall conclusions in the performance of each model. However one can compare the geographical results of the models by calculating the areal extents of each models identified clusters and overlaying the results. From the extent to which the clusters from each model overlay, calculations could be made to quantify the prevalence of the identified clusters between models. This research only employed a few of the many cluster detecting models and only one Individual Level model; additional models may act to support or discredit these findings and should be attempted.

There are many facets to disease cluster analysis that go beyond the information found at the case-location level. Aspects of the environment such as land cover and land use need to be looked at. Mosquito breeding micro-environments need to be identified to determine points where the vector is initially located. Also, the availability of birds and humans from these locations should be known. Examining the breeding cycle of the vector would help in identifying possible blooms in mosquito activity. In addition knowing the flying range of the vector would identify areas where humans and birds would be at risk. This research did not examine the roosting patterns of the bird population; this information would identify the range of the host, and combined with the mosquito information would identify the host population at risk. Research into the case fatality rate of WNV on the bird population, as well as the effects of the virus on the movement, and the length of time before mortality would also assist in the temporal study. One of the large drawbacks to the integrity of the dead bird data was that there was not 100% testing of the reported birds; complete testing of the birds would remove any

false positive errors in the sample. Also obtaining a GPS (Global Positioning System) location of the dead bird would remove the location error associated with address matching and the error in the base GIS layers. Examining the methods used to inform the public of the collection program could be used to see if the clusters were in areas where there was more advertising or greater access to available resources for reporting.

There are many ways this problem could be further examined. However, inaccuracies in the spatial location of the data points needs to be known when using this or similar data. By obtaining additional information about the collection program, the vector and the host, and advancing the cluster detection methods, the spatial and temporal locations of WNV hotspots can be more accurately identified and contained.

Appendix 1

Moran's I: 100 X 100 Grid Results

Grid 100		All Records	April	May	April-May
Results	Moran's I	0.322736	-0.000352	0.001741	0.001238
	E[I]	-0.000166	-0.00166	-0.000166	-0.000166
	Alpha Level	0.05	0.05	0.05	0.05
Normal Assumptions	Variance	0.000043	0.000043	0.000043	0.000043
	z-Score	49.40718	-0.028457	0.291864	0.214839
	Significance	0	0.977297	0.77039	0.829893
Random Assumptions	Variance	0.000042	0.000021	0.000037	0.000038
	z-score	49.555182	-0.040227	0.312904	0.229132
	Significance	0	0.967912	0.754354	0.818767
	s0	46762	46762	46762	46762
	s1	93524	93524	93524	93524
	s2	1472408	1472408	1472408	1472408
	b2	38.869679	3007.500332	784.596166	729.88598
Monte Carlo	Test Statistic	0.322736	-0.000352	0.001741	0.001238
	Simulations	100	100	100	100
	Regions ID	6019	6019	6019	6019
	Avg disease Freq	0.414853	0.000332281	0.0053165	0.00564878
	P-Value	0.0198	0	0.41584	0.00980229
Grid 100		June	June 1-15	June 16-30	July
Results	Moran's I	0.001531	0.001257	-0.004419	0.140594
	E[I]	-0.000166	-0.000166	-0.000166	-0.000166
	Alpha Level	0.05	0.05	0.05	0.05
Normal Assumptions	Variance	0.000043	0.000043	0.000043	0.000043
	z-Score	0.259644	0.217692	-0.650694	21.53769
	Significance	0.795138	0.827669	0.515244	0
Random Assumptions	Variance	0.000014	0.000008	0.000041	0.000042
	z-score	0.459967	0.518998	-0.663835	21.793114
	Significance	0.64554	0.603762	0.506796	0
	s0	46762	46762	46762	46762
	s1	93524	93524	93524	93524
	s2	1472408	1472408	1472408	1472408
	b2	4100.8201	4959.093115	238.764171	143.150902
Monte Carlo	Test Statistic	0.001531	0.001257	-0.004419	0.140594
	Simulations	100	100	100	100
	Regions ID	6019	6019	6019	6019
	Avg disease Freq	0.0098023	0.00564878	0.00415351	0.0657917
	P-Value	0.21782	0.13861	0	0.0198

Moran's I: 100 X 100 Grid Results

Grid 100		July 1-10	July 11-20	July 21-30	August
Results	Moran's I	0.01857	0.040511	0.117664	0.298348
	E[I]	-0.000166	-0.000166	-0.000166	-0.000166
	Alpha Level	0.05	0.05	0.05	0.05
Normal Assumptions	Variance	0.000043	0.000043	0.000043	0.000043
	z-Score	2.86688	6.223976	18.029254	45.675611
	Significance	0.004145	0	0	0
Random Assumptions	Variance	0.000039	0.000042	0.000041	0.000043
	z-score	2.993201	6.301259	18.437971	45.789031
	Significance	0.002761	0	0	0
	s0	46762	46762	46762	46762
	s1	93524	93524	93524	93524
	s2	1472408	1472408	1472408	1472408
	b2	499.920214	149.620081	266.678957	32.75664
Monte Carlo	Test Statistic	0.01857	0.040511	0.117664	0.298348
	Simulations	100	100	100	100
	Regions ID	6019	6019	6019	6019
	Avg disease Freq	0.00332281	0.0179432	0.0445257	0.303373
	P-Value	0.11881	0.0198	0.0198	0.0198
Grid 100		Aug 1-10	Aug 11-20	Aug 21-31	September
Results	Moran's I	0.239675	0.135355	0.128081	0.030819
	E[I]	-0.000166	-0.000166	-0.000166	-0.000166
	Alpha Level	0.05	0.05	0.05	0.05
Normal Assumptions	Variance	0.000043	0.000043	0.000043	0.000043
	z-Score	36.698072	20.736123	19.623081	4.741091
	Significance	0	0	0	0.000002
Random Assumptions	Variance	0.000042	0.000042	0.000042	0.000039
	z-score	36.932174	20.84748	19.706321	4.973168
	Significance	0	0	0	0.000001
	s0	46762	46762	46762	46762
	s1	93524	93524	93524	93524
	s2	1472408	1472408	1472408	1472408
	b2	79.001848	67.077472	53.700051	551.217315
Monte Carlo	Test Statistic	0.239675	0.135355	0.128081	0.030819
	Simulations	100	100	100	100
	Regions ID	6019	6019	6019	6019
	Avg disease Freq	0.145373	0.0765908	0.0814089	0.0275793
	P-Value	0.0198	0.0198	0.0198	0.0198

Moran's I: 100 X 100 Grid Results

Grid 100		October
Results	Moran's I	-0.002824
	E[I]	-0.000166
	Alpha Level	0.05
Normal Assumptions	Variance	0.000043
	z-Score	-0.406642
	Significance	0.684271
Random Assumptions	Variance	0.00004
	z-score	-0.419803
	Significance	0.674629
	s0	46762
	s1	93524
	s2	1472408
	b2	374.190165
Monte Carlo	Test Statistic	-0.002824
	Simulations	100
	Regions ID	6019
	Avg disease Freq	0.00265825
	P-Value	0

Moran's I: 10 X 10 Grid Results

Grid 10		All Records	April	May	April-May
Results	Moran's I	0.638524	-0.038401	0.277096	0.320102
	E[I]	-0.012821	-0.012821	-0.012821	-0.012821
	Alpha Level	0.05	0.05	0.05	0.05
Normal Assumptions	Variance	0.003544	0.003544	0.003544	0.003544
	z-Score	10.941305	-0.429699	4.870024	5.592438
	Significance	0	0.667415	0.000001	0
Random Assumptions	Variance	0.003436	0.001915	0.003309	0.003331
	z-score	11.111575	-0.584623	5.039594	5.768703
	Significance	0	0.558801	0	0
	s0	510	510	510	510
	s1	1020	1020	1020	1020
	s2	14272	14272	14272	14272
	b2	5.213742	37.525974	7.904171	7.45371
Monte Carlo	Test Statistic	0.638524	-0.038401	0.277096	0.320102
	Simulations	100	100	100	100
	Regions ID	79	79	79	79
	Avg disease Freq	31.6076	0.0253165	0.405063	0.43038
	P-Value	0.0198	0	0.01198	0.0198
Grid 10		June	June 1-15	June 16-30	July
Results	Moran's I	0.13602	0.046154	0.225586	0.527694
	E[I]	-0.012821	-0.012821	-0.012821	-0.012821
	Alpha Level	0.05	0.05	0.05	0.05
Normal Assumptions	Variance	0.003544	0.003544	0.003544	0.003544
	z-Score	2.500229	0.990652	4.004762	9.079569
	Significance	0.012411	0.321856	0.000062	0
Random Assumptions	Variance	0.001272	0.000831	0.003261	0.003334
	z-score	4.173752	2.045626	4.174575	9.360875
	Significance	0.00003	0.040793	0.00003	0
	s0	510	510	510	510
	s1	1020	1020	1020	1020
	s2	14272	14272	14272	14272
	b2	51.176094	60.531952	8.92303	7.38014
Monte Carlo	Test Statistic	0.13602	0.046154	0.225586	0.527694
	Simulations	100	100	100	100
	Regions ID	79	79	79	79
	Avg disease Freq	0.746835	0.43038	0.316456	5.01266
	P-Value	0.0198	0.11881	0.0198	0.0198

Moran's I: 10 X 10 Grid Results

Grid 10		July 1-10	July 11-20	July 21-30	August
Results	Moran's I	0.031894	0.503241	0.50607	0.616699
	E[I]	-0.012821	-0.012821	-0.012821	-0.012821
	Alpha Level	0.05	0.05	0.05	0.05
Normal Assumptions	Variance	0.003544	0.003544	0.003544	0.003544
	z-Score	0.751116	8.668813	8.716343	10.574689
	Significance	0.452583	0	0	0
Random Assumptions	Variance	0.002959	0.003266	0.003306	0.003411
	z-score	0.822005	9.030203	9.024867	10.77906
	Significance	0.411074	0	0	0
	s0	510	510	510	510
	s1	1020	1020	1020	1020
	s2	14272	14272	14272	14272
	b2	15.345473	8.828025	7.982491	5.751674
Monte Carlo	Test Statistic	0.031894	0.503241	0.50607	0.616699
	Simulations	100	100	100	100
	Regions ID	79	79	79	79
	Avg disease Freq	0.253165	1.36709	3.39241	23.1139
	P-Value	0.33663	0.0198	0.0198	0.0198
Grid 10		Aug 1-10	Aug 11-20	Aug 21-31	September
Results	Moran's I	0.640958	0.456352	0.38235	0.148966
	E[I]	-0.012821	-0.012821	-0.012821	-0.012821
	Alpha Level	0.05	0.05	0.05	0.05
Normal Assumptions	Variance	0.003544	0.003544	0.003544	0.003544
	z-Score	10.982188	7.881169	6.638089	2.717691
	Significance	0	0	0	0.006574
Random Assumptions	Variance	0.003421	0.003181	0.003374	0.003254
	z-score	11.177209	8.318739	6.803161	2.836359
	Significance	0	0	0	0.004563
	s0	510	510	510	510
	s1	1020	1020	1020	1020
	s2	14272	14272	14272	14272
	b2	5.528253	10.633845	6.532733	9.090461
Monte Carlo	Test Statistic	0.640958	0.456352	0.38235	0.148966
	Simulations	100	100	100	100
	Regions ID	79	79	79	79
	Avg disease Freq	11.0759	5.83544	6.20253	2.10127
	P-Value	0.0198	0.0198	0.0198	0.0198

Moran's I: 10 X 10 Grid Results

Grid 10		October
Results	Moran's I	0.039855
	E[I]	-0.012821
	Alpha Level	0.05
Normal Assumptions	Variance	0.003544
	z-Score	0.884852
	Significance	0.376236
Random Assumptions	Variance	0.003339
	z-score	0.911544
	Significance	0.362009
	s0	510
	s1	1020
	s2	14272
	b2	7.267707
Monte Carlo	Test Statistic	0.039855
	Simulations	100
	Regions ID	79
	Avg disease Freq	0.202532
	P-Value	0.39604

Appendix 2

Oden's I(pop): 10 X 10 Grid Results

Grid 10		All Records	April	May	April-May
Results	Ipop	8.80E-05	-7.11E-07	6.05E-07	6.56E-07
	Ipop'	0.142425	1.43784	0.0764277	0.077993
	E[I]	-2.47E-07	-2.47E-07	-2.47E-07	-2.47E-07
	Alpha Level	0.05	0.05	0.05	0.05
	% within	48.302394	147.504496	89.833083	87.297819
	% among	51.697606	-47.504496	10.166917	12.702181
Approximation	Variance	1.18E-12	1.18E-12	1.18E-12	1.18E-12
	z-score	81.3225	-0.427705	0.785711	0.832702
	significance	0	0.668866	0.432037	0.405013
Randomization	Variance	1.15E-12	5.79E-13	1.12E-12	1.12E-12
	z-score	82.2297	-0.609419	0.806939	0.854395
	significance	0	0.542247	0.419702	0.392886
Monte Carlo	Test Statistic	0.142425	-1.43784	0.0764277	0.077993
	Simulations	100	100	100	100
	Upper Tail	0.0099	0.9703	0.18812	0.18812
Grid 10		June	June 1-15	June 16-30	July
Results	Ipop	7.40E-06	8.38E-06	-2.98E-07	2.60E-05
	Ipop'	0.507311	0.997917	-0.0482642	0.265002
	E[I]	-2.47E-07	-2.47E-07	-2.47E-07	-2.47E-07
	Alpha Level	0.05	0.05	0.05	0.05
	% within	90.900942	105.537876	94.374259	66.35111
	% among	9.099058	-5.537876	5.625741	33.64889
Approximation	Variance	1.18E-12	1.18E-12	1.18E-12	1.18E-12
	z-score	7.05319	7.95783	-0.0471115	24.1575
	significance	0	0	0.962424	0
Randomization	Variance	1.13E-12	1.12E-12	1.11E-12	1.15E-12
	z-score	7.19124	8.16514	-0.0486028	24.4528
	significance	0	0	0.961236	0
Monte Carlo	Test Statistic	0.507311	0.997017	-0.0482642	0.265002
	Simulations	100	100	100	100
	Upper Tail	0.0099	0.0099	0.44554	0.0099

Oden's I(pop): 10 X 10 Grid Results

Grid 10		July 1-10	July 11-20	July 21-30	August
Results	Ipop	4.81E-07	4.46E-06	2.35E-05	8.33E-05
	Ipop'	0.0971934	0.166938	0.354976	0.184354
	E[I]	-2.47E-07	-2.47E-07	-2.47E-07	-2.47E-07
	Alpha Level	0.05	0.05	0.05	0.05
	% within	137.130233	73.822841	63.28211	45.809247
	% among	-37.130233	26.177159	36.71789	54.190753
Approximation	Variance	1.18E-12	1.18E-12	1.18E-12	1.18E-12
	z-score	0.671283	4.33921	21.9211	76.9889
	significance	0.50204	0.000014	0	0
Randomization	Variance	1.09E-12	1.14E-12	1.15E-12	1.15E-12
	z-score	0.696141	4.40705	22.2024	77.8534
	significance	0.48634	0.00001	0	0
Monte Carlo	Test Statistic	0.0971934	0.166938	0.354976	0.184354
	Simulations	100	100	100	100
	Upper Tail	0.16832	0.0099	0.0099	0.0099
Grid 10		Aug 1-10	Aug 11-20	Aug 21-31	September
Results	Ipop	6.72E-05	2.58E-05	4.06E-05	2.01E-05
	Ipop'	0.310556	0.226085	0.334904	0.490066
	E[I]	-2.47E-07	-2.47E-07	-2.47E-07	-2.47E-07
	Alpha Level	0.05	0.05	0.05	0.05
	% within	45.532464	55.756862	55.326296	56.988446
	% among	54.467536	44.243138	44.673704	43.011554
Approximation	Variance	1.18E-12	1.18E-12	1.18E-12	1.18E-12
	z-score	62.1914	23.9943	37.648	18.7783
	significance	0	0	0	0
Randomization	Variance	1.15E-12	1.15E-12	1.15E-12	1.14E-12
	z-score	62.9083	24.2833	38.099	19.041
	significance	0	0	0	0
Monte Carlo	Test Statistic	0.310556	0.226085	0.334904	0.490066
	Simulations	100	100	100	100
	Upper Tail	0.0099	0.0099	0.0099	0.0099

Oden's I(pop): 10 X 10 Grid Results

Grid 10		October
Results	Ipop	8.69E-06
	Ipop'	2.19463
	E[I]	-2.47E-07
	Alpha Level	0.05
	% within	65.135795
	% among	34.864205
Approximation	Variance	1.18E-12
	z-score	8.235
	significance	0
Randomization	Variance	1.08E-12
	z-score	8.59629
	significance	0
Monte Carlo	Test Statistic	2.19463
	Simulations	100
	Upper Tail	0.0099

Works Cited

Allen, T. R. and B. Shellito (2004). "Spatial Interpolation and Prediction of Mosquito Vectors for Surveillance of West Nile Virus." Geocarto International **19**(2).

Anderson, J. F., C. R. Vossbrinck, et al. (2001). "A Phylogenetic Approach to Following West Nile Virus in Connecticut." Proceedings of the National Academy of Sciences of the United States of America **98**(23): 12885-12889.

Anselin, L. (1995). "Local indicators of spatial association-LISA." Geographical Analysis **27**: 93-115.

Bailey, T. C. and A. C. Gatrell (1995). Interactive Spatial Data Analysis, Longman.

Caliper (2006). TransCAD. Newton, MA.

Chambers, J. and e. al (2006). The S System, Bell Laboratories.

Cliff, A. D. and J. D. Ord (1981). Spatial Processes, Models and Applications, Pion, London.

Craven, R. B. and J. T. Roehrig (2001). "West Nile Virus." Journal of the American Medical Association **286**(6): 651-653.

Dodd, R. Y. (2003). "Emerging Infections, Transfusion Safety, and Epidemiology." The New England Journal of Medicine **349**(13): 1205.

Dohm, D. J., M. L. O'Guinn, et al. (2002). "Effect of environmental Temperature on the Ability of *Culex pipiens* (Diptera: Culicidae) to Transmit West Nile Virus." Journal of Medical Entomology **39**(1): 221-225.

Eidson, M., L. Kramer, et al. (2001). "Dead bird surveillance as an early warning system for West Nile virus." Emerging Infectious Diseases **7**: 631-635.

Enserink, M. (2000). "The Enigma of West Nile Virus." Science **290**: 1482-1484.

Fortin, M.-J. (1999). Spatial statistics in landscape ecology. Landscape Ecological Analysis: Issues and Applications. New York, Springer: 253-279.

Gatrell, A. C., T. C. Bailey, et al. (1996). "Spatial Point Pattern Analysis and Its Application in Geographical Epidemiology." Transactions of the Institute of British Geographers **21**(1): 256-274.

Huhn, G. D., J. J. Sejvar, et al. (2003). "West Nile Virus in the United States: An Update on an Emerging Infectious Disease." American Family Physician **68**(4): 653-660.

Jelinski, D. and J. Wu (1996). "The modifiable areal unit problem and implications for landscape ecology." Landscape Ecology **11**: 129-140.

Klassen, A. C., M. Kulldorff, et al. (2005). "Geographic clustering of prostate cancer grade and stage at diagnosis, before and after adjustments for risk factors." International Journal of Health Geographics **4**(1).

Komar, N. (2000). "West Nile viral encephalitis." Revue Scientifique et Technique **19**: 166-176.

Komar, N., S. Langevin, et al. (2003). "Experimental infection of North American birds with New York 1999 strain of West Nile viral." Emerging Infectious Diseases **9**: 311-322.

Kulldorff, M. (1997). "A Spatial scan statistic." Communications in Statistics **26**: 1481-1496.

Kulldorff, M. (1999). Spatial scan Statistics: models, calculations, and applications, in Scan Statistics and Applications. Boston, Birkhauser.

Kulldorff, M. and N. Nagarwalla (1995). "Spatial disease clusters: detection and inference." Statistics in Medicine **14**: 799-810.

Lawson, A. B. and K. Kleinman (2005). Spatial & Syndromic Surveillance for Public Health, Wiley.

Lewin-Koh, N. J. and R. Bivand (2006). maptools: Tools for reading and handling spatial objects.

Longley, P. A., M. F. Goodchild, et al. (2001). Geographic Information Systems and Science, John Wiley & Sons Ltd.

Marra, P. P., S. Griffing, et al. (2004). "West Nile Virus and Wildlife." BioScience **54**(5): 393-402.

McLean, R., S. Ubico, et al. (2001). "West Nile virus transmission and ecology in birds." Annals of the New York Academy of Sciences **951**: 54-57.

Meade, M. S. and R. J. Earickson (2000). Medical Geography Second Edition, Guilford Publications.

Moran, P. A. P. (1950). "Notes on continuous stochastic phenomena." Biometrika **37**: 17-23.

Nychka, D. (2005). fields: Tools for spatial data.

Oden, N. (1995). "Adjusting Moran's I for population density." Statistics in Medicine **14**: 17-26.

Openshaw, S. (1984). The Modifiable Areal Unit Problem. Norwich, CATMOG: 38-41.

Openshaw, S. (1995). "Developing automated and smart spatial pattern exploration tools for geographical information systems applications." the Statistician **44**(1): 3 - 16.

Openshaw, S., M. Charlton, et al. (1988). "Investigation of Leukaemia Clusters by use of a Geographical Analysis Machine." The Lancet **331**(8580): 272-273.

Pebesma, E. J. (2006). gstat: geostatistical modelling, predicting and simulation.

Petersen, L. R. and A. A. Marfin (2002). "West Nile Virus: A Primer for the Clinician." Annals of Internal Medicine **137**(3): 173-179.

Plante, M., K. Lowell, et al. (2004). "Studing deer habitat on Anticosti Island, Quebec: relating animal occurrences and forest map information." Ecological Modelling **174**: 387-399.

Qi, Y. and J. Wu (1996). "Effects of changing spatial resolution on the results of landscape pattern analysis using spatial autocorrelation indicies." Landscape Ecology **11**: 39-49.

Rowlingson, B. S. and P. J. Diggle (2006). splancs: Spatial and Space-Time Point Pattern Analysis.

Team, R. D. C. (2006). R: A Language and Environment for Statistical Computing. R. F. f. S. Computing. Vienna, Austria.

Tsai, T. F., F. Popovici, et al. (1998). "West Nile encephalitis epidemic in southeastern Romania." The Lancet **352**: 767-771.