Spatial and temporal analysis of the West Nile Virus in the United States with special emphasis on Louisiana

by

Verena Huber

2nd Bachelor Thesis

Submitted in partial fulfillment of the requirement of the degree
Bachelor of Science

Carinthia University of Applied Science
School of Geoinformation

Supervisors

Supervisor: Dr. Michael Leitner

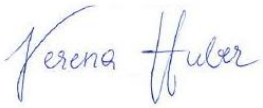
Department of Geography and Anthropology, Louisiana State University,

Baton Rouge, USA

Villach, June 2012

Science Pledge

By my signature below, I certify that my thesis is entirely the result of my own work. I have cited all sources I have used in my thesis and I have always indicated their origin.



Villach, 15.06.2012

Verena Huber

Acknowledgements

I would like to give my special thanks to my supervisor Dr. Michael Leitner, who made this thesis project possible in the first place. Thank you for your help and advice and the chance, you gave me, to spend a semester at the Louisiana State University. Also, thank you for having read my thesis, chapter for chapter and thank you for the valuable feedback.

A special thank you goes to my family. Their understanding and endless support helped me a lot throughout my education. Thank you for making it possible to take this academic degree, and thank you for accompanying and supporting me on that way. I am grateful for all the help and advice I got from you.

Zusammenfassung

Das West-Nil-Virus ist eine Infektionskrankheit, welche ihren Ursprung in den späten Dreißigerjahren im West-Nil-Distrikt in Uganda hat. Auf Menschen wird das Virus mittels eines Moskitobisses übertragen. Nach einigen verzeichneten Virusausbrüchen in Europe trat das Virus erstmals 1999 in den USA zu Tage (CDC, 2011a). In den Folgejahren verbreitete sich das Virus über den gesamten nordamerikanischen Kontinent und verursachte in manchen Fällen schwere Erkrankungen, wie zum Beispiel eine Entzündung des Gehirns (Enzephalitis), oder eine Entzündung der Hirn- und Rückenmarkshäute (Meningitis). Laut des CDC verursachte das WNV seit seinem ersten Auftreten im Jahr 1999 in den USA mehr als 1200 Todesfälle (CDC, 2011a). In den folgenden Jahren verbreitete sich das Virus nicht nur über die gesamten Vereinigten Staaten, sondern es erschien auch in Kanada und einigen Staaten in Südamerika. Abhängig von einigen Erfolgsfaktoren taucht das Virus in bestimmten Regionen auf, verursacht Vogelsterben und einige Krankheitsfälle unter Menschen und Pferden, bevor sein Auftreten schwächer wird und es schließlich wieder ganz verschwindet. Die Forschung dieser Bachelorarbeit befasst sich im Wesentlichen damit, Verteilungsmuster des WNV in den USA zu untersuchen. Dabei wird der Zeitraum vom ersten Auftreten im Jahr 1999 bis 2011 untersucht. Die Analysen werden mit Daten über Krankheitsfälle, welche vom CDC zur Verfügung gestellt werden, durchgeführt. Die Krankheitsfälle werden zu Analysezwecken zu den administrativen Einheiten (Staaten) aggregiert. Zusätzlich werden auf einer größeren Maßstabsebene Analysen für den Staat Louisiana durchgeführt. Daten über WNV-Fälle pro Parish (administrative Einheiten in Louisiana, sind äquivalent zu den „Counties“ in anderen Staaten oder den Bezirken in Österreich) werden vom Louisiana Department of Health and Hospitals zur Verfügung gestellt. Der Grundnutzen dieser Arbeit besteht darin, so genannte „Cluster“, das sind nahe bei einander liegende Gruppierungen von Daten, über Zeit und Raum aufzudecken. Dies dient dem Zweck, die Verteilungsmuster des WNV besser zu verstehen. Zudem geben Cluster, welche in der Vergangenheit aufgetreten sind, Aufschluss über mögliche Verteilungsmuster und Clusterbildungen in der Zukunft. Für die Raum-Zeit-Analyse von WNV-Fällen werden unterschiedliche Methoden verwendet: der „local indicator for spatial association“ (LISA), retrospektive und prospektive Tests, die Kulldorff Scan Statistik und die Visualisierung von multivariaten Raum-Zeit-Muster mit Hilfe einer Selbstorganisierenden Karte (SOM). Die statistischen Analysen werden in einer geographischen Informationssystem (GIS)-Umgebung durchgeführt. Die Analysen werden mit frei erhältlichen statistischen Software-Paketen durchgeführt. Dazu gehören: Open GeoDa (Anselin et al., 2004), GeoSurveillance (Rogerson et al., 2009), SaTScan (Kulldorff, 2011), und Vis-Stamp (Guo, 2006a). Diese Pakete unterscheiden sich stark in ihren Funktionalitäten und ihrer Software-Architektur. Einige von den Programmen wurden noch nie im Zusammenhang mit Gesundheitsdaten oder im Speziellen mit Daten über Krankheitsfälle verwendet. Deshalb wird im Zuge dieser Bachelorarbeit auch die Anwendbarkeit dieser Programme für Gesundheitsdaten überprüft. Die Ergebnisse aus den Analysen können durch Evaluierungen und einem Vergleich mit anderen Ergebnissen validiert werden. Die meisten Softwarepakete verfügen über keine ausreichenden Kartierungs- beziehungsweise Visualisierungsfunktionen. Daher wird für diese Zwecke mit der kommerziellen GIS-Software ArcGIS von ESRI (ESRI, 2011) gearbeitet.

Abstract

The West Nile Virus (WNV) is an infectious disease which has its origins in the late thirties in the West Nile District of Uganda. Humans acquire the virus through a mosquito bite. After some reported human outbreaks of the infection in Europe the virus first entered the U.S. in 1999 (CDC, 2011a). In the following years it spread over the North American continent and in some cases it caused severe illness such as inflammation of the brain (encephalitis) or inflammation of the spinal cord (meningitis). According to the Centers of Disease Control and Prevention (CDC) since its first occurrence in 1999 the virus-borne West Nile Disease caused over 1200 deaths in the U.S. alone (CDC, 2011a). In subsequent years the virus has not only spread over the U.S. but also entered Canada and several states in South America. Depending on different factors, the virus usually emerges in a region causing avian die-offs and several disease cases among humans and horses before it ceases again. The research in this Bachelor thesis aims to analyze spatio-temporal WNV distribution patterns across the U.S. from its beginnings in 1999 until 2011. The analysis is carried out on count data provided by the CDC. Data have been subsequently aggregated to administrative units (states). In addition, analysis is carried out on a larger scale for the State of Louisiana. Data about WNV incidents on a parish level in Louisiana are provided by the Louisiana Department of Health and Hospitals. The main purpose of the analyses is to detect clusters across space and time. This will help to better understand the distribution patterns of the WNV. Furthermore, clusters which have emerged in the past give an idea of possible future distribution patterns of WNV incidents. For the spatial and temporal analysis of WNV incidents several techniques are implemented, including a local indicator for spatial association (LISA), retrospective and prospective tests of clusters and clustering, the Kulldorff's Scan Statistic, and the visualization of multivariate space-time patterns using a self-organizing map (SOM). The statistical analyses are conducted in a Geographic Information System (GIS) environment. Principal components of this environment are freely available statistical software packages, including Open GeoDa (Anselin et al., 2004), GeoSurveillance (Rogerson et al., 2009), SaTScan (Kulldorff, 2011), and Vis-Stamp (Guo, 2006). These packages differ significantly in their functionalities and software architecture. Some of them have never before been applied to health data, in general or disease data, in particular. Thus, the software packages' applicability for health data is tested in the course of this Bachelor thesis. After evaluating and comparing outputs of the analyses, respective results can be validated. Most software packages used in this thesis lack functionalities for mapping and visualizing results. This makes the use of a GIS inevitable. For this purpose, the commercial ArcGIS software from ESRI (ESRI, 2011) is used.

Table of Contents

List of Figures

List of Tables

List of Formulas

List of Abbreviations

CDC	Center for Disease Control and Prevention
CSR	Complete Spatial Randomness
DIANA	Divisive Analysis
EB	Empirical Bayes
ESDA	Exploratory Spatial Data Analysis
ESRI	Environmental Systems Research Institute
FIPS	Federal Information Processing Standard
GIS	Geographic Information System
GR	Geary's Ratio
LISA	Local Moran's I spatial Autocorrelation or Local Index of Spatial Association
MAUP	Modifiable Areal Unit Problem
MC	Moran Coefficient
PCP	Parallel Coordinate Plot
PEP	Population Estimates Program
PVD	Presumptive Viremic Donors
SOM	Self Organizing Map
SSCM	Sum of Squares Clustering Model
TIGER	Topologically Integrated Geographic Encoding Reference
WNV	West Nile Virus

1 Introduction

Alexander the Great suffered from a 2-weeks febrile illness terminating in flaccid paralysis and encephalopathy. The emperor died of the illness in the ancient Mesopotamian city of Babylon on June 10, 323 BC. The reason for his sudden passing away remains a controversial issue among medical investigators. Recent theories state that Alexander the Great may have died of the West Nile Virus (WNV) encephalitis (Marr & Calisher, 2003). A factor, which brings scientists to this conclusion, is that simultaneously to his illness, bird observers which were common at that time observed an inexplicable sudden avian die-off. Especially, endemic ravens were affected by the epizootic (Marr & Calisher, 2003). But only recently, as the WNV became a more global issue, it was discovered that there might be a connection between the avian die-off and Alexander's fatal illness. Originally, the WNV was endemic to Asia, Africa, Europe as well as Australia (Petersen, 2009). Over the ages it has emerged in some areas and disappeared in others. Across time the virus does not only move geographically, but it also underlies lots of mutation processes. Due to that fact, the WNV might appear in a more aggressive form for avian species in an area, affecting susceptible species and causing avian die-offs accompanied by a few human febrile cases. Depending on the viremia, WNV might also cause severe human disease cases, including WNV encephalitis and WNV meningitis before the virus ceases again. In 1999, it emerged in the Western hemisphere for the first time. It then spread rapidly across the entire American continent. In 2002, the virus triggered the largest arboviral (mosquito-borne) encephalitis epidemic in US history (Huhn et al., 2003). This is when the virus reached an unprecedented prevalence among human beings. The conditions which led to that peak are broadly unknown.

Since the first emergence of the WNV in the U.S., the Center of Disease Control and Prevention has created a national passive surveillance system (ArboNET) to collect both human and non-human data about incidences of arboviral infections, such as a WNV infection (CDC, 2009a). The objectives of this system are to monitor incidence as well as geographic and temporal spread of WNV and other arboviruses. Additionally, the system should be a source of information for public health officials. The system is fed by local health departments and it is updated on a weekly basis (CDC, 2009a). The Louisiana Health Department provides data about WNV activity on a parish-level. They issue annual reports about WNV activity in the state.

Geographic information systems provide an adequate environment to visualize disease transmission patterns over space. In disease modeling the temporal dimension also plays a primary role. However, the integration of a third, the temporal dimension is still a challenge for developers in the GIS field. Since space-time analyses have become a more and more important issue in many disciplines, such as epidemiology, a few solutions to include the temporal component in a GIS have been proposed through the open source and freeware community. The latter community is also a striving new trend in modern information technology. Freely available software comes along with many possibilities but also with a few risks. For free developers there exists neither a requirement list nor any criteria catalogue for developing software products. Especially, when it comes to data input and output, freeware products provide very different solutions.

Research papers about the analysis of WNV incidences using GIS are available on a wide range. But space-time investigations with statistical tools are rare. Wimberly, Lindquist, and

Wey have analyzed the equine WNV outbreak in South Dakota in 2002 using the commercial ArcGIS by ESRI (ESRI, 2011) and Anselin's free GeoDa (Anselin et al., 2006) for spatial autocorrelation analysis (Wimberley et al., 2011). Other research uses GIS analytical tools to integrate risk factors and create a WNV risk map for humans. This was implemented by Rochlin et al. in 2011 in the paper "Predictive Mapping of Human Risk for West Nile Virus (WNV) based on Environmental and Socioeconomic Factors" (Rochlin et al., 2011). In this context an analysis about GIS-supported WNV risk modeling on the Mississippi river has been conducted. This study reveals high human risk zones along the Mississippi river (Cooke et al., 2006). Another highly interesting research area is the investigation of an association between different land cover areas and WNV risk. Ruiz et al. have investigated the association of West Nile virus illness and urban landscapes in Chicago and Detroit (Ruiz et al., 2007). The purpose of that writing is to improve the understanding of human exposure to WNV-infected mosquitoes in an urban context (Ruiz et al., 2007). Again, GIS was used to integrate various environmental factors. A similar research was conducted by Bowden, Magori, and Drake who investigated how human disease and land cover types are associated across the U.S. (Bowden et al., 2011).

The main objective of this Bachelor Thesis research is to examine behavior of transmission patterns and their circumstances year by year and across the US, based on county-level data. In detail, transmission patterns will be investigated for the state of Louisiana. The research questions for this Bachelor Thesis are as follows:

- Is it possible to conduct space-time analysis with data collected from the CDC Website as well as on from the Website of the Louisiana Health Department with freely available software programs?
- Do the free statistic software packages provide satisfactory tools for exploratory data analysis, cluster detection, and visualization of results?
- Is it possible to visualize multivariate WNV patterns across space and time?
- How do WNV patterns change over time and vary across space in the U.S and Louisiana?
- How do temporal trends differ in different places compared to places with WNV activity?
- Does the trend observed during the observation period allow a short term prediction of future human WNV infections?

The aim of this Bachelor Thesis is to conduct a space-time analysis with disease data and freely available software programs. The disease cases are aggregated to administrative units, such as states or counties. Important methods that will be introduced in this thesis are methods for the analysis of outliers, the global and local Moran's I spatial autocorrelation, retrospective analysis over space and time for historic data, prospective analysis for the investigation of future trends and Kulldorff's scan statistic. The analyses are conducted in order to detect space-time clusters and a clustering effect over space and time. The methods discussed in this Bachelor Thesis should be of interest to researchers of various fields who conduct space-time analysis with areal summarized data, where individual point data are expressed either as a count of events or a rate.

This Bachelor Thesis is organized as follows: The first chapter gives a brief introduction into the research. The second chapter will give an introduction into the epidemiology and

ecology of the WNV. The focus in this chapter will be on geographic distribution patterns and how WNV emerged in the United States and eventually in Louisiana. In addition, there will be an introduction to the virology of WNV and the effect it has on humans.

The third chapter starts with a description of the data used in this research. Additionally, the free software packages used for space-time analysis are introduced. Selected cluster analysis techniques, integrated in the software will be explained in the last section of the third chapter.

The fourth chapter contains the results of the space-time analysis of WNV in the U.S. and in Louisiana. This chapter includes various thematic maps, graphs and charts to visualize analysis results.

The last and fifth chapter contains the conclusion about the analysis results as well as prospective research fields.

2 Epidemiology

West Nile Virus (WNV) was first isolated from the blood of a febrile woman in the West Nile district of Uganda in 1937 (Petersen, 2009). It is considered to be the earliest arthropod-borne virus discovered by humans (White & Morse, 2001). Prior to 1996 several epidemics had been documented in rural areas. Most infections were asymptomatic and only a few cases of severe neurological disease had been reported. WNV remained an occasional cause of febrile illness in Africa, the Middle East, parts of Europe and Russia, South Asia and Australia. For this reason WNV was not considered to be a significant human pathogen (White & Morse, 2001). Between 1996 and 1999, however, the virus unexpectedly triggered major epidemic activity in southern Romania, the Volga delta in southern Russia and the northeastern United States, involving hundreds of cases of severe neurological disease and fatal infections. This was also the first time when WNV-borne epidemics reached urban populations (White & Morse, 2001). The severe epidemics in the 1990s had some circumstances in common. The common house mosquito *Culex pipiens* was apparently involved as vector in all three cases. As a matter of fact all three urban areas in Romania, Russia and the United States had lower than normal rainfall during the summers of epidemics. This condition could have been responsible for an increase of potential breeding sites for *Culex pipiens*. All three areas were located adjacent to large rivers, which provide an adequate habitat for resident and migratory species of wild birds (White & Morse, 2001). Additionally, the epidemic in the northeastern United States was accompanied by epizootic in birds, especially corvids in this area (Scheld et al., 2007).The high virulence to American crows from the NY99 WNV strain was caused by a single nucleotide change in the viral DNA (Petersen, 2009). In Romania 400 neuroinvasive cases of WNV infection have been reported in the time range from July 15 until October 12, 1996 (White & Morse, 2001). The peak of the epidemic activity was in early September. The epidemic was geographically confined to fifteen districts in the Danube plain of southeastern Romania including the urban area of Bucharest. There are a few factors which contributed to the epidemic in Romania. One of them was the poverty and deteriorated suburban respectively urban infrastructure that resulted in abundant *Culex pipiens pipiens* larval habitat. There was little precipitation in spring and summer and it was a very hot summer which promoted the production of *Cx. p*.

pipiens. Furthermore, the human population in the fifteen districts was highly susceptible to WNV infection. Prior to the outbreak of the epidemic serum samples were tested for antibodies to WNV. The samples from Bucharest were all negative to antibody, making people highly vulnerable to a new emerging WNV epidemic. After the incident in Romania case studies were conducted. Risk factors for WNV infection were the presence of mosquitoes in the home, more mosquito bites per day, flooded basements of apartments, and spending a greater amount of time outdoors (White & Morse, 2001).

2.1 Geographic distribution patterns

Since its first discovery in 1937 in Uganda, the WNV has spread considerably all over the planet. The distribution of the virus depends on climate, appropriate mosquito breeding sites, and bird habitats. These influencing factors have resulted in certain distribution patterns. The epidemiology follows several patterns (Petersen, 2009). There is widespread enzootic transmission throughout tropical Africa, the Caribbean, Central America and northern South America without significant human or equine morbidity. In the Mediterranean Basin, Russia and South Africa there happen to be periodic human and equine outbreaks followed by low-level enzootic activity periods. In India incidents are reported sporadically. In Southeast and East Asia there is little WNV enzootic activity and no human cases. Australia faces from time to time small outbreaks and sporadically human disease cases. In North America repeated annual outbreaks are reported during the summer time.

2.2 West Nile Virus in the United States

The virus was first detected in North America during a human outbreak of meningitis and encephalitis in 1999 where five people in Queens were hospitalized due to severe illness (Petersen, 2009). Simultaneously epizootic activity in birds, especially in corvids was reported (Scheld et al., 2007). Tests revealed that the viral DNA was 99.8% identical to a WNV strain of an Israeli goose (White & Morse, 2001). For this reason the origin of the North American outbreak might lie in the Mediterranean basin. Until then it was not known that the virus can be fatal for birds. The high virulence to American crows from the NY99 WNV strain was caused by a single nucleotide change in the viral DNA (Petersen, 2009). This change is likely to be the reason why the virus became lethal to some birds. Prior to the 1999 outbreak no WNV cases have been recorded in the Western Hemisphere. The origin of the outbreak is unknown. Possible origins could be an infected bird (migrated or imported), a viremic individual, or an infected mosquito (White & Morse, 2001). An additional favorable circumstance for the outbreak was the abnormally hot weather.

In the following years the WNV spread extensively towards the west. In 2002 and 2003 multistate outbreaks in the Midwestern states resulted in more than 2,800 reported neuroinvasive cases each year. From 2004 to 2007 reports revealed a lower number of WNV-borne neuroinvasive diseases (Petersen, 2009) (see Figure 2.1). Human infection rates are increasing in the period between April and October. They reach a peak in August or early September. Risk factors for a WNV infection are farming, vegetation abundance in urban area and living in an inner suburb (Petersen, 2009).

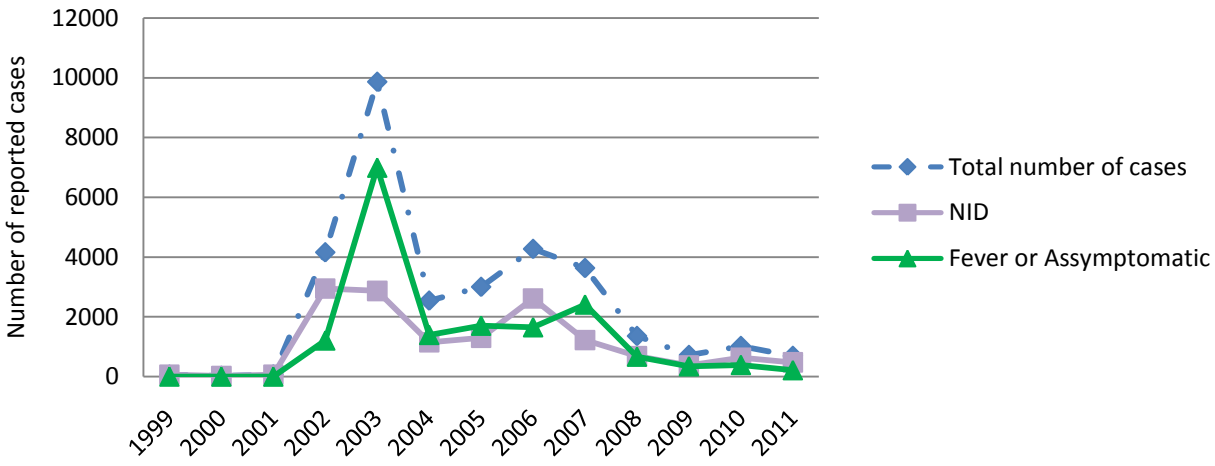


Figure 2.1: Human WNV-infections in the U.S. 1999-2011

2.3 West Nile Virus in Louisiana

In 2001 the WNV appeared for the first time in Louisiana in the Parishes Vermillion, Jefferson, Plaquemine, Calcasieu, and Iberia. Six birds, one human, and ten horses tested positive for the virus (Gruszynski, 2006). In 2002 the virus spread more aggressively. There were 329 human cases in 31 parishes. The WNV activity was found nearly in all 64 parishes of the state. There have been 17 fatalities among the human cases. More than 80% of those cases belonged to the age group over 60 (DHH, 2002). Statistically, Louisiana had the second highest incidence of confirmed human WNV cases in the nation in 2002. It ranked first in the number of mortalities due to the WNV (Gruszynski, 2006). In 2003 human and equine WNV positive cases decreased slightly. According to the Louisiana Health Department there were 122 human cases in 32 parishes (DHH, 2003). In the following years there were slight fluctuations of incident cases among human but the 2002 peak has not been surpassed till today (see Figure 2.2).

In Louisiana the main vector of the WNV is *Culex quinquefasciatus*, the common southern house mosquito. A study in St. Tammany in 2002 revealed the potential of several species of birds found in Louisiana as amplifying hosts. Major hosts are northern cardinals, house sparrows, blue jays and northern mocking birds. Apart from humans and horses as incidental hosts also alligators tested positive for WNV in an investigation in 2003 (Gruszynski, 2006).

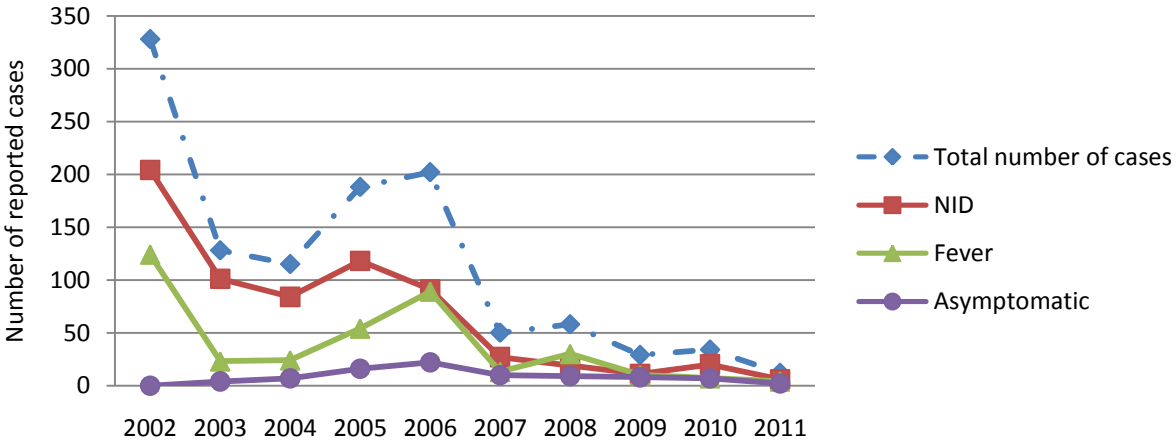


Figure 2.2: Human WNV-infections in the Louisiana 2002-2011

3 Virology

Arboviruses or arthropod-borne viruses are defined as viruses that require a hematophagous (blood-sucking) arthropod for transmission. These roles are usually played by mosquitoes or ticks which are part of a complex transmission cycle. Apart from the arthropod (mosquito) the cycle also involves a primary vertebrate reservoir host. Mostly birds or rodents are primary hosts. They develop enough viremia in their bodies that mosquitoes feeding on them will be infected with the virus. Humans or domestic animals are usually so called incidental or dead-end hosts. They do not produce enough viremia to contribute to the transmission cycle.

3.1 Human Arthropod-borne virus infections

Currently, there are 534 viruses registered. Of these, 134 types have caused illness in human (White & Morse, 2001). Three families form the group of the arboviruses: *Bunyaviridae*, *Flaviviridae* and *Togaviridae*. Viruses of these families are the most important human pathogens as far as public health is concerned. The WNV is a single-stranded RNA virus of the family of *Flaviviridae*, genus *Flavivirus*.

Arboviruses are distributed all over the world. In the last few centuries a few new types of viruses have been discovered. More important is, however, that currently some silently behaving already known viruses experience resurgence and extensive geographic spread. The reasons for this revival might be societal changes and the modern transportation network. The distribution and development of arboviruses, however, depends on a few limiting factors or ecologic parameters. These are temperature, precipitation and vegetation patterns. These parameters directly influence the success of arthroprods and vertebrate hosts. The better the conditions are for those two species, e.g. higher temperatures and lower precipitation in summer, the better are the chances for the virus to thrive and move on to new areas. Especially in times of climate change as temperature and precipitation patterns change the virus emerges in new non-endemic areas. Additional factors that are responsible for the resurgence of arboviruses are a global population growth, movement of people within and among regions, and changes in agriculture. Also the pathogens itself have been changing, so that there is an increased movement of viruses in humans and animals. Genetic change will lead to an increased potential of epidemics (White & Morse, 2001).

3.2 The virus

Shortly after the original isolation of the WNV in Uganda, researchers found out that the WNV was antigenically related to two other arboviruses that were known to cause encephalitis, namely the St. Louis encephalitis (SLE) virus and the Japanese encephalitis (JE) virus (White & Morse, 2001). That means that the isolates of the viruses as well as their stems are serologically widely identical (Wordnik, 2012). Later, further studies expanded the relationship of the WN virus with many other flaviviruses like Murray Valey encephalitis, Kunjin, Usutu, Kokobera, Stratford, and Alfuy viruses. Thus, the WNV was assigned to the Japanese encephalitis virus serocomplex (Petersen & Marfin, 2002), which is a group of antigenically closely related, mosquito-borne flaviviruses that are responsible for severe encephalitic disease in humans (Lobigs et al., 2009). With new and improved DNA-sequencing technologies the WN isolates could be divided genetically into two lineages: I and II. Major human outbreaks of the WNV have been associated only with lineage I WNVs. Lineage II WNVs are maintained in enzootic cycles primarily in Africa and are not associated

with human or animal outbreaks (Petersen, 2009). Improved technologies allowed researches to determine the relationship of WNV isolates made during the course of the three urban epidemics between 1996 and 1999. The testing showed that all three isolates belonged to lineage I of the WNV. All of the isolates had a high degree of homology (>99.8%) (White & Morse, 2001).

3.3 Transmission cycles

The WNV has two distinct transmission cycles (see Figure 3.1). There is a primary enczootic or amplification cycle involving one set of vectors and avian hosts and secondary cycles involving potentially different arthropods and transmission to other hosts such as humans and horses (White & Morse, 2001). In the primary cycle of the WNV, ornithophilic mosquitoes, such as genera from the *Culex*, feed on viremic birds (amplification hosts). They become infected and are capable of transmitting the WNV to other amplification hosts. Birds are also called reservoir hosts. They develop high enough titers of virus (viremia) to infect mosquitoes, if they feed on them. If environmental conditions such as temperature, mosquito species, mosquito population density, and a number of susceptible hosts are given, an epizootic will occur in the avian population. An epizootic in birds, however, will not necessarily result in human or equine disease. Primary vectors feed exclusively on avian hosts and especially on those species which develop high-level viremias. Due to this fact they are highly efficient amplification vectors but do not pose high risk of transmitting the WNV to humans. On the other side mosquitoes that are more general feeders are no efficient vectors but they could be a much greater threat to humans and equines. This species are known as bridge vectors. They could become infected when feeding on an infected bird and then transmit the WNV to a susceptible vertebrate host (White & Morse, 2001). Humans and domestic animals are incidental hosts; they do not produce high enough viremia-levels to contribute significantly to the transmission cycle.

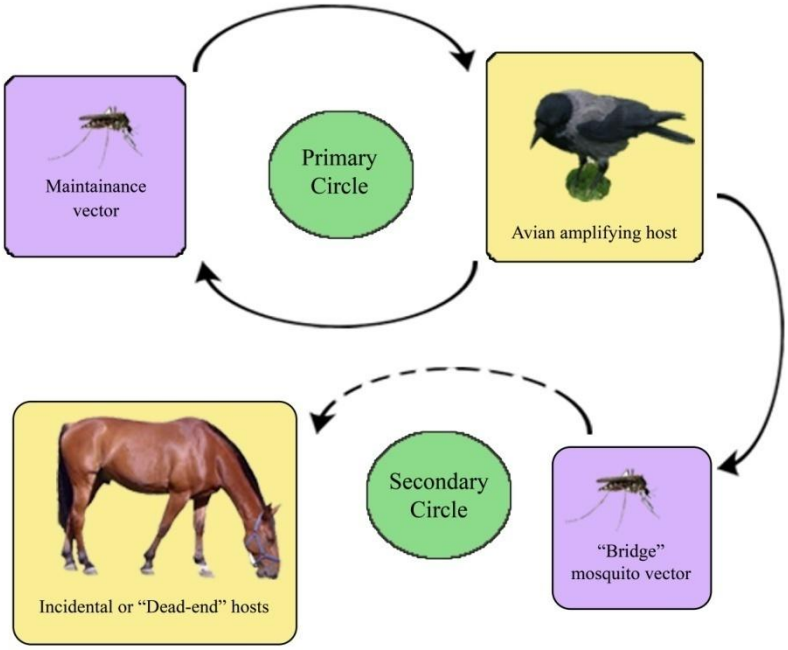


Figure 3.1: The WNV transmission circle

3.4 The vector

Studies in the early 1950s revealed that several species of mosquitoes can be infected and successfully transmit the WNV. However, over the years isolates have primarily been made from pools of mosquitoes belonging to the *Culex* family (White & Morse, 2001). Subgenera of the *Culex* depending on the area were identified to be the most important vector in the WNV primary transmission cycle. Several species from the family of the *Aedes* or *Ochlerotatus* have been implicated as bridge vectors (White & Morse, 2001). The WNV has also been isolated occasionally from ticks. Their role is still undefined but they could be important for overwintering the WN virus in temperate areas. In 1999 researchers observed that infected females of the family of *Culex pipiens pipiens* were able to overwinter, transporting the virus to the next season (White & Morse, 2001).

The *Culex* mosquito, or the common house mosquito is one of the major types of mosquitoes inhabiting the planet (Tiny Mosquito, 2012). It typically obtains its blood meal from birds instead of humans. Due to this fact the *Culex* is not considered to be that harmful to humans than other mosquito families like Anopheles and *Aenes* are. The *Culex* likes to lay her eggs on the surface of standing fresh or stagnant water. It prefers outdoor objects on people's property such as barrels, cans and garden pots to plant, and wildlife surroundings (Mosquito-Netting, 2012). The female *Culex* deposits between 100 and 300 eggs onto the water surface. Two days later the larvae will hatch. Once the larvae have hatched they will stay seven to fourteen days beneath the water surface. With the help of a siphon the larvae can grasp oxygen from the surface. In this time span the larvae goes through four developing stages. Thereafter, the mosquito larvae become a pupa. This stage will last one to four days. Afterwards, the adult mosquito breaks through the pupa, rests until its body has dried and hardened completely and then flies away to find a partner for pairing. Female *Culex* mosquitoes attack vertebrate hosts, preferable birds to blood-feed after dawn (Mosquito-Netting, 2012). If they feed on a WNV infected bird with high enough virus titers, they also become infected and are able to transmit the WNV to other vertebrate animals. Infected mosquitoes which were born in late autumn can overwinter and conserve the virus till the next season. In southern states of the U.S. mosquitoes are active all year round.

According to Petersen in the United States more than 60 mosquito species have been infected with the WNV (Petersen, 2009). The genus of mosquito known as the *Culex* acts as vector in the transmission cycle. Depending on the geographic area a certain subgenus of *Culex* is responsible for the transmission: *Cx. pipiens* and *Cx. restuans* in the northern U.S. and Canada, *Cx. quinquefasciatus* in the southern U.S. and *Cx. tarsalis* in the western U.S (Petersen, 2009).

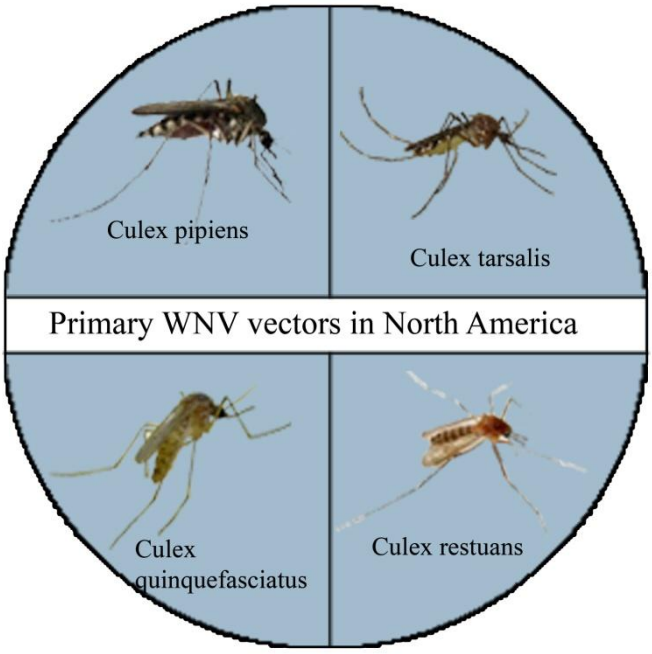


Figure 3.2: Primary WNV vectors in North America

3.5 The host

Humans were the first known vertebrate hosts of the WNV. This was proven by virus isolations from the blood as well as the presence of the WNV neutralizing antibodies in the blood (White & Morse, 2001). Once infected, humans as well as domestic animals do not develop high enough titers of virus in their blood to efficiently infect other mosquito vectors. This is why humans and most of the domestic animals are so called incidental or dead-end hosts. They do not play a crucial role in the transmission cycle and are only involved accidentally. By contrast, studies show that in the blood of advanced cancer patients, titers of the WNV are high enough as to potentially infect mosquitoes on blood-feeding (White & Morse, 2001). Unlike incidental hosts, amplifying hosts play an important role in the transmission cycle. Amplifying hosts or reservoir hosts develop high titers of the virus and are capable of infecting mosquitoes. A study conducted in Egypt in 1952 showed that wild birds are important amplifying hosts in the transmission (White & Morse, 2001). The study revealed that wild birds develop enough viremia in their blood. Additionally, the birds had high rates of antibodies to the WNV in their blood. Another aspect is that high WNV activity among humans goes along with high WNV activity in birds. When talking about the epidemic in New York in 1999, there was a simultaneous epizootic in birds, which makes the transmission cycle complete. In the course of antibody surveys the WNV has been isolated or identified serologically in many native and imported vertebrate species in North America. These include: bats, wolves, eastern foxes, gray squirrels, chipmunks, sheep, alligators, alpacas, black bears, macaques, reindeers, dogs, monkeys and baboons, raccoons, skunks, and opossums (Petersen, 2009). Their roles in the transmission of WNV are, however, undefined or they fall into the group of incidental hosts.

3.6 Alternative ways of transmission

The majority of human WNV infections results from mosquito bites. However, there are several alternative routes of how the WNV can be transmitted from human to human. One direct way is via blood transfusion. This was proven in the U.S. outbreak in 2002 when 23 blood recipients became infected after receipt of platelets, red blood cells or fresh frozen

plasma from viremic blood donors (Petersen, 2009). Other ways of transmission are organ transplantation, transplacental transmission, breast milk transmission, and dialysis-related transmission (Scheld et al., 2007).

3.7 West Nile Virus and humans

Once infected, typical incubation period ranges from 2 to 14 days. Longer incubation periods have been observed among the immune-suppressed (Petersen, 2009). Approximately 20 – 30% of persons develop illness after infection with lineage I strains (Petersen, 2009). The remaining 70 – 80% does not experience any symptoms and stays asymptomatic. Due to this fact most human infections are not clinically apparent. Many infections remain undetected, since most affected persons do not see a physician. Clinical disease ranges from mild febrile illness to severe encephalitis. Based on their clinical presentation, arboviral disease cases are often categorized into two primary groups: Neuroinvasive disease and non-neuroinvasive disease (CDC, 2011b), which is also known as West Nile fever. Advanced age is by far the most significant risk factor for severe neurologic disease. In the 1999 New York epidemic the incidence of severe neurologic disease was ten times higher in persons 50 to 59 years of age compared to persons younger than 19 (Petersen & Marfin, 2002). A further independent risk factor for neuroinvasive disease is diabetis mellitus, which was found out during the 1999 New York City outbreak (Scheld et al., 2007).

The diagnosis of the WNV in a person rests on a high index of clinical suspicion and on results of specific laboratory tests. The WNV as well as other arboviral diseases should be seriously considered in elderly adults who experience a sudden onset of unexplained encephalitis or meningitis in summer or fall (Petersen & Marfin, 2002). However, in southern states of the U.S. transmission can occur year-round. In addition, the local prevalence of the WNV enzootic activity should also raise suspicion. The most efficient diagnostic method is the detection of Immunoglobulin M (IgM) antibody to WNV in serum or cerebrospinal fluid. Since IgM antibody does not cross the blood-brain barrier, IgM antibody in cerebrospinal fluid strongly suggests central nervous system infection (Petersen & Marfin, 2002).

3.7.1 Non-neuroinvasive disease

The usual clinical presentation is called the West Nile fever, which was coined by Goldblum in 1952 (Scheld et al., 2007). The clinical symptoms of the disease are: an acute onset of fever, severe frontal headache, malaise, back pain, myalgias, general weakness, drowsiness, anorexia and fatigue (CDC, 2011b; Scheld et al., 2007). Eye pain, pharyngitis, nausea, vomiting, diarrhea, abdominal pain and rash can also occur (Petersen, 2009). The rash associated with the WN fever is characterized by a flat, red area on the skin that is covered with small confluent bumps and it predominates over the torso and extremities, sparing palms, and soles (Scheld et al., 2007). Clinical diagnosis for a non-neuroinvasive infection defined by the Center of Disease Control and prevention (CDC) is fever (>100.3°F or 38°C) reported by the patient or a health-care provider, absence of neuroinvasive disease and absence of more likely clinical explanation (CDC, 2011b). The acute illness lasts three to six days. However, convalescence is slow and can range from one to two weeks, which is accompanied by general fatigue (Scheld et al., 2007).

3.7.2 Neuroinvasive disease

Neuroinvasive disease occurs approximately in 1 of 140 infected persons. Clinical manifestation includes encephalitis, aseptic meningitis, or flaccid paralysis (Petersen, 2009).

These illnesses are usually characterized by the acute onset of fever with stiff neck, altered mental status, seizures, limb weakness, cerebrospinal fluid (CDF) pleocytosis that is an increase of white blood cells in the CDF, or abnormal neuroimaging (CDC, 2011b). Additionally, the course of the disease can be accompanied by tremor, myoclonus (involuntary twitching of muscels) and Parkinsonian features such as rigidity, postural instability, and bradykinesia. The clinical diagnosis for a neuroinvasive infection defined by the CDC is fever (>100.3°F or 38°C) reported by the patient or a health-care provider and meningitis, encephalitis, acute flaccid paralysis, or other acute signs of central or peripheral neurologic dysfunction, as documented by a physician, and the absence of a more likely clinical explanation.

The WNV-borne meningitis is clinically similar to any other viral meningitis. Affected persons experience the abrupt onset of fever, headache, nuchal rigidity, photophobia, or phonophobia. In general, WNV meningitis is associated with a favorable outcome. In the 2002 U.S. epidemic 2% of all registered WNV meningitis cases were fatal (Scheld et al., 2007). WNV encephalitis can range from a mild, self-limited confusional state to severe encephalopathy, coma, and death. Clinical symptoms are generally associated with movement disorders. This is due to a specific neurotropism of WNV for regions of the brain involved with control of movement (Scheld et al., 2007). Tropism is generally the ability of a virus to infect a certain type of cell or tissue (Modrow et al., 2003). It appears that the WNV has a predilection for neurons in the central nervous system (Scheld et al., 2007). Movement disorders manifest in coarse tremor, myoclonus and Parkinsonism. Primary tremor and Parkinsonism may persist in patients recovering from severe encephalitis. Long-term effects of the WNV encephalitis can include persistent neurologic dysfunction and movement disorders, brain damage and permanent muscle weakness (Modrow et al., 2003; The New York Times, 2010). The WNV can also cause paralysis. This might happen if viremia damages the lower motor neurons of the spinal cord, resulting in acute flaccid paralysis (Modrow et al., 2003). Fatality rates among patients with severe neuroinvasive disease range from 10 to 20% (Modrow et al., 2003).

4 Methodology

The following section provides information about the data and software used in the analysis. Furthermore, important clustering techniques are introduced.

4.1 Data

For the analysis it is necessary to have data that include information about time, disease incidences, and geographic locations. Both data sets for the U.S. and Louisiana are available annually on a sub-division-level. For the U.S. the subdivisions are states and for Louisiana the subdivisions are parishes, which are equivalent to counties. This information needs to be aggregated based on geographic location. For this purpose, polygon shapefiles from the U.S. Census Bureau can be used. The final input data must be compliant with the software programs in use. In certain analysis the software GeoDa does not support so-called island and doughnut polygons. Island polygons do not have any neighbors whereas doughnut polygons are completely surrounded by one or many other polygons. This had to be taken into consideration before starting to actually run the software with the input data.

Additionally, the data has to be somewhat modified for input into each new software package. Details about the data adaption will be given in the analysis section.

4.1.1 WNV in the United States

Data about WNV incidents in the U.S. are documented by the CDC. Since the first appearance of WNV in the U.S. the CDC has initiated an electronic surveillance program to document WNV activity. The program is called ArboNET (https://idepi.oph.dhh.la.gov/ArboNet). It is a dynamic system collecting both human and non-human data with the purpose to monitor incidence as well as geographic and temporal spread of WNV and other arboviruses. It provides information for public health officials, government officials, and the public. The content can be divided into two major categories: Ecologic data and human data. Ecologic data are for example veterinary cases (horses), dead birds, mosquitoes and sentinel like for example chicken. Human cases can be distinguished between neuroinvasive cases, non-neuroinvasive cases or presumptive viremic donors (PVDs) (CDC, 2009a). PVDs are people with no symptoms when donating blood. However, their blood tests positive in preliminary tests, when screening blood for the presence of WNV (CDC, 2011a). Blood screening has become more and more important after the outbreak in the U.S. in 1999 in order to reduce alternative ways of WNV transmission. Data from ArboNET is provided for free on a state-level.

4.1.2 WNV in Louisiana

In Louisiana surveillance of arboviral activity is conducted by the center for community and preventive health, which is a department of health and hospitals of the state of Louisiana (http://new.dhh.louisiana.gov/index.cfm/page/539). The surveillance program for WNV was initiated in spring 2000, a year after the first emergence of WNV in the Western hemisphere. Currently the program involves testing of dead and live birds, sick horses, mosquito pools, and sentinel chicken flocks (DHH, 2012). An annual report about WNV activity was first available in 2002. The reports include observed WNV activity on parish-level or divided into age groups as far as human cases are concerned. Additionally, the report provides temporal and geographical statistics about bird cases, equine cases and mosquito pools infected with WNV. The human cases were initially divided into meningo-encephalitis (ME), fever, unknown, and fatalities. However, the terminology has changed in 2004. Now, the categories are neuroinvasive disease (NID), fever, asymptomatic cases, and fatalities (DHH, 2012).

4.1.3 Census Data

In this project census data on a state level are required in order to create WNV infection rates. These rates are visualized in form of choropleth maps. Furthermore, more detailed population data about Louisiana is needed to display WNV incidence on a parish level.

Current data about the population of each U.S. state are provided by the census bureau. The U.S. census bureau conducts a population census every ten years. The last one took place in 2010. The data derived from the census are used to determine the number of seats each state has in the U.S. House of Representatives. They are also used to distribute funds to local communities (U.S. Census, 2010). The time frame regarded in this research ranges from the first appearance of WNV in the United States in 1999 to 2011. This period includes two population censuses which can directly be used for the calculation of raw rates. In order to take annual population changes into account estimated population data have been used.

These estimated data are provided in a summarized form by infoplease.com and are available for the years 2004, 2006, 2007 and 2008 (Infoplease, 2012). For the remaining years, with no timely population estimates available, data from existing estimations were used. For example, WNV data for the years 1999-2002 are based on 2000 population census data. WNV data for 2003-2005, the 2004 population estimates were used. The WNV data analysis from 2009-2011 are based on the 2010 population census data.

Population data on parish level for the state of Louisiana are also provided by the U.S. census bureau. Parishes are political subdivisions of Louisiana and analogous to counties in other states (Louisiana.gov, 2012a).The 2010 census was compared with data from the 2000 census. There was a population decrease in 25 of the 64 parishes. In seven parishes there was an increase of more than 15% (Louisiana.gov, 2012b). WNV had its first onset in Louisiana in 2001. Thus, population data are required for analysis for the period from 2001 to 2011. Apart from the census data of 2000 and 2011 the U.S. census bureau provides annual population estimates for the state of Louisiana. The Census Bureau's Population Estimates Program (PEP) is responsible for population estimates. It utilizes current data on births, deaths, and migration to calculate population change since the most recent decennial census and produces a time series of estimates of population, demographic components of change, and housing units. The estimates are produced on a national, state and county level (U.S. Census Bureau, 2012).

Base maps of the U.S. as well as for the state of Louisiana are also available at the U.S. Census Bureau. The shapefiles can be downloaded from the TIGER (Topologically integrated geographic encoding reference) database. The files provide the digital map base for a GIS or mapping software. All legal boundaries and names are as of January 1, 2011 (U.S. Census Bureau, 2011a). The download of TIGER files is a two-step process. First, a layer type has to be determined. This can be an administrative division like, for example, counties and states or a feature type such as roads, water, and railways. Then the search engine will display the geographic areas for which the chosen layer type is available. The layers underlie specific name conventions. The first part of the name is reserved for tl_2011, which is the abbreviation for tiger line and the year it was created. The next letters indicate the geographic extent. If the geographic extent is the entire nation, then the abbreviation is "us". If it is state or county-based the unique Federal Information Processing Standard (FIPS) code is used. Further parameters are layer type and file extension. The content of the attribute table are administrative subdivisions, abbreviation of administrative subdivisions, FIPS code, amount of land area, and water area. Each shape file has a .prj-file which contains projection information. All Census Bureau generated shapefiles are in Global Coordinate System North American Datum of 1983 (GCS NAD83) (U.S. Census Bureau, 2011b). A big advantage of TIGER shapefiles is that they have already been generalized by the census bureau. This is important for analysis purposes. For example, this is necessary when spatial autocorrelation contiguity weights files have to be created. Therefore, the data should not include so called "island polygons" which would bias results and result interpretation. Thus, the type of analysis proposed in this thesis could neither be conducted for Alaska or Hawaii. These two states were thus excluded from the study area.

4.2 Software

The software environment of this thesis is freely available and includes powerful tools for temporal, spatial, and exploratory data analysis. However, the packages do not provide any

or satisfactory visualization tools so that it was necessary to use an external GIS software with mapping and visualization options. Therefore, the commercial ArcGIS, Version 10.0 from ESRI was applied for visualization purposes.

The free software packages have to fulfill several criteria in order to be adequate for the analysis conducted within this thesis. The most important criterion is that the software is freely available and downloadable from the internet. At the same time, it has to work in a Microsoft Windows operating system and it has to be under active development. The program should also be alone-standing so that no programming skills are necessary to run it (Anselin, 2003a). Eventually, the program should come with a user's guide and a technical documentation. Ideally, there are some tutorials and sample data which are specially edited for the program in order to get started with the software environment.

The main purpose of the spatial and temporal analysis is the detection of clusters and clustering over space and time. According to Anselin a software environment should include several essential requirements to carry out an exploratory analysis of disease clusters (Anselin, 2003a). Cluster analysis software should have an efficient interface to a GIS, in the sense of providing means to extract the relevant data and to feed back results for map display. There must be effective data input, meaning that the program is able to read in x and y coordinates of locations for cases. In case of areal aggregated data the software has to be capable of processing digital boundaries of polygons containing information about events, rates, or risk estimates. Additionally, the statistical software program should be able to construct spatial weights either on a distance or contiguity base. It should provide means for spatial autocorrelation analysis like the global and local Moran's I spatial autocorrelation (LISA). Furthermore, since distance and quadrat tests are common methods for cluster detections the program should provide these tools, too. For visualization purposes and a better understanding of the results there should be options to create maps and graphs. Eventually, the program should process a flexible program output. Thus, the results can be integrated with other software, such as a GIS (Anselin, 2003a). Four programs have been chosen for the analysis and evaluation. Among other things, the purpose of this thesis is to find out, whether they are suitable for disease data analysis (see Table 4.1).

Name	Version	Date (yyyy-mm-dd)	Author and Institution
Open GeoDa	1.0.1	2011-10-20	Luc Anselin Center for Spatially Integrated Social Science at the University of Illinois, Urbana-Champaign http://geodacenter.asu.edu/
GeoSurveillance	1.1	2007-06-06	Gyoungju Lee, Ikuho Yamada, and Peter A. Rogerson NCGIA (National Center for Geographic Information and Analysis), Department of Geography, State University of New York at Buffalo http://www.acsu.buffalo.edu/~rogerson/geosurv.htm
VIS-Stamp	1.0	2009-06-17	Diansheng Guo Department of Geography, University of South Carolina http://www.spatialdatamining.org/software/visstamp
SaTScan	9.1.1	2011-03-09	Martin Kuldorff Harvard Medical School, Boston, and Information Management Services Inc, Silver Spring, Maryland http://www.satscan.org/

Table 4.1: Free software packages for spatial and temporal data analysis

4.2.1 Open GeoDA

Open GeoDA is a free software package developed at the University of Illinois and Arizona State University. It is a collection of software tools designed to implement techniques for exploratory spatial data analysis (ESDA) (Anselin, 2003b). The main purpose of GeoDA was to create a software package where several windows with different views of the same data can be linked dynamically. When data are highlighted in one of the active views the same observations should also be highlighted in all the other views (Eck et al., 2005). An extension to linking several views in order to get a dynamic framework is brushing. Thereby a rectangle is created with the pointer which can be moved in the graphs. The selected observations change depending on where the rectangle is moved and this affects all current windows (Anselin, 2003c). The functionalities of GeoDA can be divided into six categories. These are: Spatial data manipulation, data transformation with the possibility of creating new variables, ESDA, mapping data, spatial autocorrelation, and spatial regression (Anselin et al., 2004). In this project GeoDA is used for ESDA, mapping data, and spatial autocorrelation analysis. In the category mapping GeoDA provides tools to create choropleth maps, cartograms, and map animations. Map animations are short map movies, demonstrating changes in a pattern. GeoDa provides several options to display the spatial distribution of rates. Creating choropleth maps always goes along with a number of challenges. The underlying risk is the inherent variance instability or unequal precision of rates. Therefore, GeoDa contains five mapping routines to deal with the visualization of rates in maps. These are: Raw rate, excess risk, empirical bayes, spatial rate, and spatial empirical bayes (Anselin, 2003b).

Spatial autocorrelation in GeoDa is defined by means of the Moran's I spatial autocorrelation statistic and the visualization in form of a Moran Scatter Plot, which is a special type of a scatter plot (Anselin, 2003b). There are two different types of spatial autocorrelation. The first one is global spatial autocorrelation, which implements tests for clustering (Anselin, 2003c). Therefore a Moran Scatter Plot is created containing the value of the Moran's I which reveals positive, negative, or no spatial autocorrelation among the data. The second one is local spatial autocorrelation. The local spatial autocorrelation analysis helps to identify clusters in data sets and determines the level of significance of clusters. This statistic is essential for creating LISA (Local Moran's I spatial autocorrelation) maps. The cluster detection analysis with spatial autocorrelation in GeoDa is based on a weights file. This file can either be a distance weights file where the distances between points, such as X and Y coordinates are taken into consideration or a contiguity weights file. The latter contains neighbor information of the polygons in the area of investigation. The choice is between Rook-based or Queen-based contiguity. Data which were processed with different contiguity weights can differ because Rook contiguity uses only common boundaries to define neighbors, while Queen Contiguity includes all common points, like boundaries and vertices (Anselin, 2003b). This is the reason why spatial weights created with the Queen option results in more neighbors than those with the Rook option.

GeoDa can read ESRI Shapefiles and dbf-files as data input. Using the field calculator, new fields can be added and values can be calculated. With the option "save table" these fields can be saved to the original table but this is not the default. GeoDa offers to export of maps to bitmap files (bmp) or to portable network graphics (png). Additionally, created rates for choropleth maps can be saved to the attribute table. This makes sense, when working with the data in ArcGIS. The results of LISA maps can also be added to the attribute table. For the

spatial correlation type an indicator value is stored, which takes on the value of 1 for high-high, 2 for low-low, 3 for low-high and 4 for high-low (Anselin, 2003b).

4.2.2 GeoSurveillance

GeoSurveillance is a small, stand-alone freeware program which combines spatial statistical routines with some basic Geographic Information System (GIS) functions (Rogerson et al., 2007). It was developed by Peter A. Rogerson, Gyoungju Lee, and Ikuho Yamada at the University of New York at Buffalo. The GIS functions are limited to map-display related operations, such as loading maps, coloring based on legend schemes, zooming in, zooming out, and panning (Rogerson et al., 2007). The software can carry out both retrospective and prospective tests for the detection and monitoring of spatial clustering. Retrospective tests are applied to spatial data collected for a particular point in time. They contain a testing of the null hypotheses of no spatial clustering. The null spatial model describes the spatial distribution of cases expected in the absence of clustering (Waller & Jacquez, 1995). A small p-value, depending on the assumed significance level, would reject the null hypothesis of the absence of spatial clustering and assume that the alternative hypothesis is true. The alternative hypothesis is defined as “not the null hypothesis”. That means that the alternative hypothesis is favored if the data is inconsistent with the null hypothesis (Waller & Jacquez, 1995). Still, there is a chance that the wrong hypothesis either the null hypothesis or the alternative hypothesis is assumed to be true. A type one error happens if the null hypothesis is rejected although it is true. In the course of a type two error the null hypothesis is assumed to be true although it is wrong. The significance level α determines the chances of the occurrence of a type one, the power of a test, β, the chance of committing a type two error.

GeoSurveillance contains a set of integrated statistical tests. For retrospective tests GeoSurveillance uses the local score statistic to determine whether incidence is raised around a predefined spatial location (Rogerson et al., 2007) as well as the spatial M statistic, which examines the maximum local score statistic (Rogerson et al., 2009). Additionally, a global score statistic gives a summary of the local statistics.

Prospective tests analyze time-series data to detect emergent clusters as quickly as possible and incorporate the dynamic nature of data (Rogerson et al., 2009). For prospective analysis GeoSurveillance applies the univariate cumulate sum (cusum) method for a normal variable to individual subregions in a study area (Rogerson et al., 2009). The cusum tool assumes normality, thus the investigated variable has to be transformed into a normal variate beforehand (Rogerson et al., 2009). Its main purpose is to detect deviations in a variable of interest, from one mean to the other. Z-scores in various forms are used as input variables in the prospective procedure (Rogerson et al., 2007). Additionally, GeoSurveillance presents a chart which highlights the highest univariate cusum value among locations.

GeoSurveillance can be divided into three major components: A cluster detection and monitoring component, a GIS component, and a support tool component (Rogerson et al., 2009). The first component can further be divided into a tool for retrospective analysis and one for prospective analysis. Before implementing a retrospective analysis it is necessary to choose the statistical type, which can be adjusted or unadjusted. For retrospective testing the score statistic, adjusted or unadjusted, and the spatial M statistic are available. The local M statistic is the spatially weighted z-score in a locality i. GeoSurveillance uses three

different ways of transforming the z-value z_i into z-values for each sub-region z_j, based on observed and expected values in the sub-regions j. The formula which describes each statistic is depicted in the field “Expression of Variables”. A user has the choice to run the analysis using a single bandwidth only or determining a range of bandwidths. The significance level α is set to 0.05 which is the default in social sciences. Alternatively, this level can be changed to 0.01. GeoSurveillance is designed primarily for data that are available for a set of sub-regions. For each subregion, observed and expected numbers of cases are assumed to be available (Rogerson et al., 2007). The result of the retrospective analysis is displayed in the map window. There is a window for both the local pattern summary and the global pattern summary. Additionally, a legend for the map can be created. The legend is based on predefined threshold values. For example, when using the local score statistic the category endpoints are defined as zero, a third and two thirds of the maximum and minimum values. When choosing the spatial M statistic threshold values are defined upon the critical value of the spatial M statistic (Rogerson et al., 2007).

Prospective tests in GeoSurveillance require a set of columns (fields) that represent z-scores over time, where each column (field) corresponds to the temporal unit for monitoring (Rogerson et al., 2007). This can be for example on an annual or monthly level. The prospective test is based on the cusum method which incorporates three parameters: t, k, and h. T is for time and k and h are threshold values. The value of h is determined in conjunction with a desired false alarm rate, which is similar to a type I error rate (i.e., significance level) in ordinary retrospective analysis (Rogerson et al., 2009). In addition, a bandwidth σ can be specified by the user. If the bandwidth is set to zero, it is assumed, that the observed z-values in all regions are independent. The results reveal the maximum cusum value for each year. The program also constructs a chart for the maximum cusum. It will also summarize the signaled region and cusum value for the last year of observation. The map and the corresponding legend have a predefined color scheme. The legend is constructed based on the threshold (h-value). The cutoff values are: 25%, 50%, 75% and 100%. If the cusum value of a subregion exceeds the h-value itself, the subregion is colored red (Rogerson et al., 2007). These regions are listed in the “Signaled region and cusum value for the year” table.

The software package is able to process ESRI shapefiles as polygon data and simple text files as point data. GeoSurveillance has two restrictions: One is that the file name must not exceed nine characters and the data must be stored on a hard drive (Rogerson et al., 2007).

4.2.3 SaTScan

SaTScan is a free software to analyze spatial, temporal and space-time data using the spatial, temporal or space-time scan statistic. It was developed by Martin Kulldorff at Harvard Medical School in Boston, in cooperation with the Information Management Services Inc. The software implies a scan statistic which is used to detect and evaluate clusters of cases in either a purely temporal, purely spatial, or space-time setting. Therefore, a window gradually scans the area of investigation across time and/or space. At each location the scanner notes the numbers of observed and expected observations inside the window. If the analysis is purely temporal the scanning window is an interval in time, if it is spatial only the window assumes the shape of a circle or an ellipse. If the analysis is both spatial and temporal the scanning tool is a cylinder with a circular or elliptic base (Kulldorff, 2010). According to Kulldorff an important characteristic of the spatial scan test is that it can both

detect the location of clusters and do inference (Kulldorff, 1997); thus, locate the geographic area of the most likely cluster as well as secondary clusters on a map.

Depending on the data it can be chosen between various probability models. The Bernoulli Model includes a boolean case variable, which represents either cases (1) or non-cases (0). It is also referred to as cases and controls. As input data the Bernoulli Model requires information about the location of cases and controls. Under the discrete Poisson model the number of cases in each location is Poisson-distributed (Kulldorff, 2010). The discrete Poisson model is especially suitable for event or count data (Kulldorff, 1997). It requires case and population counts for a set of data locations such as counties, parishes, census tracts or zip code areas. Additionally, it is necessary to provide geographical coordinates for each of the locations, e.g. centroids of polygons or county seats. The Space-Time Permutation Model requires only case data and the spatial location as well as time for each case. Background population or controls are not considered in this model. The model compares the number of observed cases in a cluster to what would have been expected if the spatial and temporal locations of all cases were independent of each other. Thus, in this model it is referred to a cluster when a specific geographical area has a higher proportion of cases in a specific time period compared to the remaining geographical areas (Kulldorff, 2010). The software also contains an exponential model for survival time data, a normal model for continuous data and a model for spatial variation in temporal trends.

The program takes any text files and some table sheets, such as .dbf or .xls, as data input. These files can be imported, whereby the input parameters have to be specified. Thus, the user has to determine the column which contains the parameter. Input parameters are ID, x and y coordinates, time, population, cases, and controls. Then, the program converts the given input to a program-specific input. There are case files (*.cas) containing information about cases, coordinate files (*.geo) with the geographic locations for the cases, population files (*.pop) with the background population in a certain geographic location, and control files (*.ctl) containing non-cases. It depends on the respective model which files are eventually required as input information. For temporal and space-time analyses the number of cases must be stratified by time. Attributes of cases, such as age or gender, may also be provided (Kulldorff, 2010).

The results of analysis include a standard text based results file in American Standard Code for Information Interchange (ASCII) format. In addition, the program generates, if desired, up to five different output files in column format. They can be generated in either ASCII or dBase format. The Standard Results File reports a summary of the data and the most likely cluster as well as secondary clusters. The optional Cluster Information File (*.col) displays each cluster in a separate row. Additional information about the cluster is given in the columns. The Location Information File (*.gis) represents the cluster data in a way that is easy to incorporate into a GIS. The Risk Estimates for Each Location File (*.rr) gives information about the relative risk for each location (Kulldorff, 2010).

The program does not provide any mapping or visualization options. For effective working with SaTScan an additional GIS environment is inevitable. However, since SaTScan implies Kulldorff's scan statistic it is an essential component in each cluster detection analysis.

4.2.4 VIS-Stamp

VIS-Stamp (Visualization of Spatial, Temporal, and Multivariate Patterns) allows users to discover interesting and unknown complex patterns among a set of multivariate data. Additionally, one can investigate results of various perspectives. VIS-Stamp represents the results in a way that supports human interpretation, analytical reasoning, and decision making (Guo, 2009). The program was developed by Diansheng Guo at the Department of Geography at the University of South Carolina in 2009. Version 1.0 was issued in 2009. Since then, no new version has been published but the program remains under steady development.

The program surface of VIS-Stamp can be divided into five main windows. There is the control window (i), where the variables which shall be included in the analysis can be specified. The variables are normalized to z-scores. Additionally, a user might give each variable a weight which will be multiplied with the z-score. A Self-Organizing-Map (SOM) (ii) is responsible for the processing of the data. It derives clusters of spatial objects based on their multivariate similarity. Therefore, the SOM uses the Euclidean distance to assess multivariate similarity between spatial objects (Guo, 2009). The SOM assigns similar clusters with similar colors from a 2D color scheme. This color scheme can be dynamically rotated or flipped by a user (Guo, 2009). The clusters can be visualized in maps (iii) or in a re-orderable space-time matrix (iv). A parallel coordinate plot (PCP) (v) is used as the legend to show the multivariate vector that each color represents. The PCP displays the variables like for example disease types on the x-axis. The y-axis represents the numeric values. Each cluster is a string in the PCP, with the same color as it has in the SOM. In order to compare characteristics of various variables across space and time data should be transformed into a percentage value.

The SOM is an artificial neural network. The cells of the network become tuned to various input patterns through an unsupervised learning process. The learning process in this network is competitive, unsupervised or self-organizing (Kohonen, 1990). The SOM has been used for tasks like pattern recognition, robotics, process control and processing of semantic information (Kohonen, 1990).

The data used for analysis should be event cases inside a certain geographic boundary (polygon). If point patterns are available they can be aggregated to an areal unit. However, Guo and Wu consider this methodology as a limitation of data, since it reduces the data resolution by using a predefined set of boundaries. Alternatively, they suggest creating a kernel density surface for each variable type as well as for each time period. Each raster pixel could then be a spatial unit in the analysis. The authors also state that this approach has its own limitations as far as the uncertainty in the interpolated data is concerned .

Each data input consists of three files. First, there is the shapefile containing information about the geographic boundaries, e.g. states, counties, or zip-codes. Then there is the attribute file in .csv-format, which must have the same name as the shapefile. Finally, the spatial, temporal, and multivariate data is provided in a separate .csv-file. This file has certain conventions concerning the order of fields (columns). Guo and Wu (2012) explain the data aggregation and preprocessing by means of a data cube (see Figure 4.1). The three dimensions in the data cube include: The spatial dimension, the temporal dimension like years, months or days, and the multivariate dimension (e.g., crime types). Each cell in the

data cube is defined with a unique combination of a spatial unit, a certain time period, and a variable type (e.g., crime type). The value of a cell is, thus, the total number of, for example, crime types in that cell

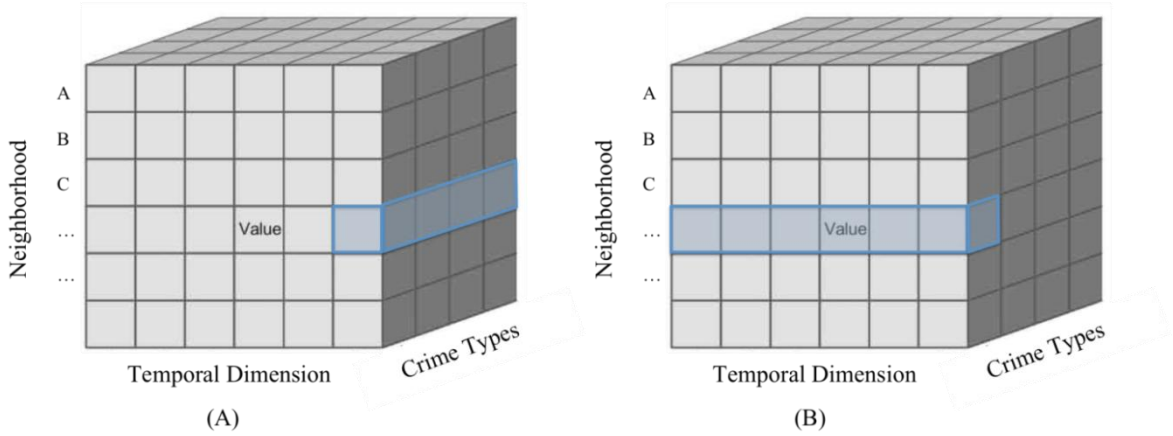


Figure 4.1: The data cube demonstrates a space-time-attribute aggregation of crime data. (A) A sequence of crime types is highlighted. (B) A time series is highlighted (taken from Guo & Wu)

4.3 Spatial clustering

4.3.1 Data types

Geocoded data about events occurring at a certain location, such as a crime incident, a tornado touchdown point, or a fatal WNV incident are generally referred to as a spatial point pattern. Patterns in a point data set can either arise through some form of clustering mechanism or through environmental variations, which lead to high concentrations of events in certain regions (Diggle, 2003). Point pattern analysis helps to understand the processes which led to a certain point pattern. In this analysis, an expected point pattern is often compared to an observed one. As expected point patterns, usually, a complete spatial random point pattern is used. Complete spatial randomness (CSR) is the null hypothesis in spatial point pattern analysis. CSR follows a Poisson distribution with certain criteria (Diggle, 2003). CSR includes several constraints: First, it implies that the intensity of events does not vary over the plane. Second, there are no interactions amongst the events (Diggle, 2003). In a random point pattern distribution any point is equally likely to occur at any location and the position of any point is not affected by the position of any other point. Apart from random distributions there are uniform and clustered distributions. In a uniform distribution every point is as far from all of its neighbors as possible. In clustered distributions many points are spatially concentrated, while large areas contain little or no points. Event data that are summarized and aggregated to a spatial unit such as a county or a zip code are called spatial aggregated areal data. Such data are often represented as a rate, with the event data being normalized with the population, or the area of the respective unit. Using normalization, different units can be compared with each other. In thematic maps, such as choropleth maps, only rates can be displayed since the use of total counts would bias interpretation. A drawback of areal aggregated data is that a certain degree of information gets lost when aggregating event data. For example, areal aggregated data do not provide information about the respective exact location of each event. This aspect, however, can also have an advantage when dealing with confidential data. Thus, areal aggregated data provide a certain protection of revealing too personal data. For this purpose, data about health issues, such as disease incidents, are mostly available only in areal aggregated data.

4.3.2 Cluster analysis

Several expressions refer to the method of creating groups of objects: Cluster analysis, segmentation analysis, taxonomy analysis, or unsupervised classification. These groups are formed in a way that objects in one cluster or group are similar, whereas objects from different clusters are distinct (Gan et al., 2007). Everitt (Xu & Wunsch, 2009) defines a cluster as follows: "A cluster is a set of entities which are alike and entities from different clusters are not alike." Xu and Wunsch (Xu & Wunsch, 2009) state, that classifying or grouping data into a set of categories or clusters plays an important and indispensable role in the history of human development. Thus, if there is an unknown object or phenomenon, people try to identify descriptive features of this object and compare these features with those of known objects or phenomena (Xu & Wunsch, 2009). Based on their similarity or dissimilarity the new object or phenomena can be categorized as related or not related. Since data capturing methods are improving significantly, more and more data become available. In order to deal with an increasing amount of data, clustering as well as classification play an important role to explore and summarize that data.

Classifications as well as clustering methods are both components of the data mining process. In data mining large amounts of data are explored and analyzed in order to retrieve useful information. Classifications belong to the subgroup called direct data mining. Thus, data items are assigned to predefined classes. When talking about clustering, the clusters are not known a priori. Clustering is indirect data mining, where the goal is to discover some relationships among all the variables (Gan et al., 2007). Xu and Wunsch (Gan et al., 2007) describe it as a method where it is "not exactly sure what clusters one is looking for".

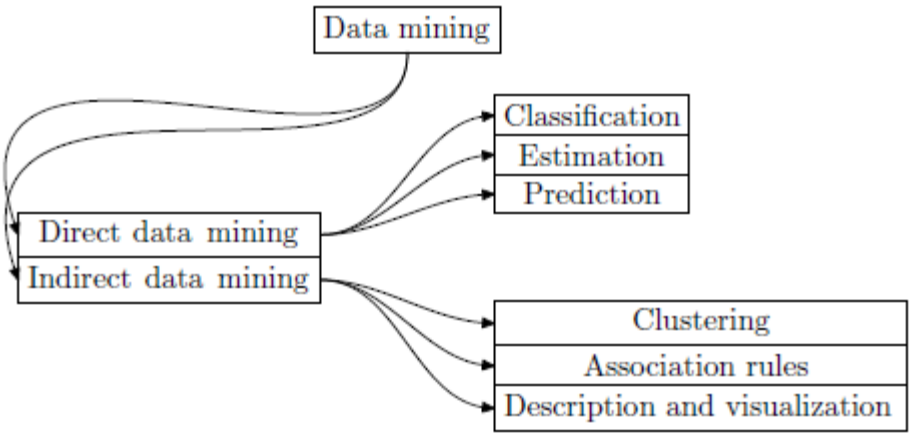


Figure 4.2: The methods of data mining (taken from Gan et al., 2007)

Objects in a cluster have to fulfill several criteria. According to Bock (Gan et al., 2007) all objects in a cluster have to share the same or closely related properties, show small mutual distances or dissimilarities, have contacts or "relations" with at least one other object in the group, or be clearly distinguishable from the rest of the objects in the data set. For numerical data there are two types of clusters: Compact clusters and chained clusters (Gan et al., 2007). The compact cluster is a set of data points with high mutual similarity, which can be displayed by a representative point or center (see Figure 4.3). In chained clusters, any member of a cluster can reach another member by following a certain path (see Figure 4.4). Apart from distinguishing clusters on basis of their appearance and structure, another method is to distinguish between hard and fuzzy clustering. Hard clustering is the idea that

each object in a data set belongs to “one and only one cluster” (Gan et al., 2007). The constraints imply that each object either belongs to a cluster or not. Furthermore, each object belongs to only one cluster and each cluster contains at least one object, so that no empty clusters are allowed (Gan et al., 2007). By contrast, there is the concept of fuzzy clustering, which assumes that an object can be member of one or more clusters. Thus, the constraints are more relaxed.

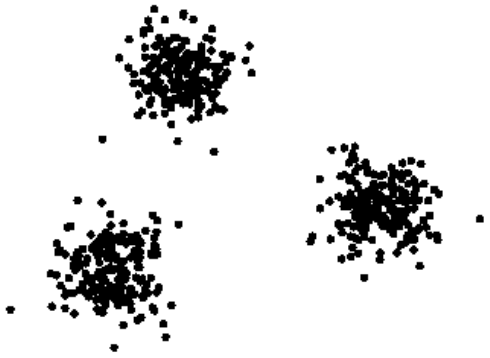


Figure 4.3: Compact clusters (taken from Gan et al., 2007)

Figure 4.4: Chained clusters (taken from Gan et al., 2007)

4.3.3 Clustering Algorithms

In clustering, distances and similarities play an important role. Distances are used to describe the similarity or dissimilarity of two data points or two clusters. In cluster analysis the Euclidean distance is created in order to develop an index of similarity (see Figure 4.1). Every clustering algorithm is based on the index of similarity or dissimilarity (Gan et al., 2007).

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{j=1}^{d} (x_j - y_j)^2 \right)^{\frac{1}{2}}$$

Formula 4.1: Calculating the Euclidean distance d between two points in a data set (x and y); (taken from Gan et al., 2007)

A clustering algorithm usually contains four design phases (see Figure 4.5): Data representation, modeling, optimization and validation (Gan et al., 2007). The first phase, data representation, gives an idea of which kind of clusters one might expect in a data set. Then, the modeling phase defines the notion of clusters and the criteria that separate different clusters. This is followed by an optimization of the quality measure. In the validation phase clustering results are assessed with validity indices. These measures are used to evaluate and assess the results of a clustering algorithm (Gan et al., 2007). Conventional clustering algorithms can be divided into two categories: Hierarchical and partitional algorithms (Xu & Wunsch, 2009). Hierarchical clustering groups data with a sequence of nested partitions, which either form singleton clusters to one large cluster including all individuals or the other way round (Xu & Wunsch, 2009). In partitional clustering data points are directly divided into some pre-specified number of clusters without any hierarchical structure (Xu & Wunsch, 2009).

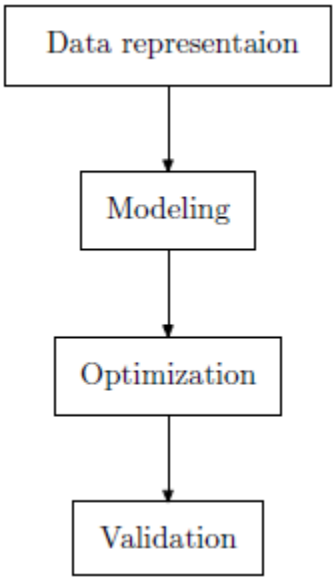


Figure 4.5: Process of data clustering (taken from Gan et al., 2007)

4.3.3.1 Hierarchical clustering

Hierarchical clustering techniques can be divided into two kinds of algorithms: There is agglomerative hierarchical clustering and divisive hierarchical clustering (see Figure 4.6). Agglomerative clustering starts with every single object in a single cluster. Then, it keeps merging the closest pairs of clusters, until all objects are forced into the same group (cluster). Thus, agglomerative clustering is considered to be bottom-up. There are graph and geometric methods in agglomerative clustering. In graph methods such as the single-link method, clusters can be represented by a sub-graph or by interconnected points. In geometric methods, such as the centroid method, a cluster can be represented by a central point. The single-link method employs the nearest neighbor distance to measure the dissimilarity between two clusters (Gan et al., 2007). In the first step of the single-link method a dissimilarity matrix is created. According to that matrix those clusters with the least distance between, are merged. This step is iterated until all data points are merged into one single cluster.

Divisive clustering starts with all objects being included in one cluster. At each step the number of clusters increases by one, since at each stage of the algorithm one cluster is divided into two (Gan et al., 2007). The algorithm keeps splitting large clusters into smaller pieces until all clusters are singletons (Xu & Wunsch, 2009). Therefore, the divisive method is also referred to as top-down algorithm. Divisive hierarchical clustering methods can be of two types: Monoethic and polyethic (Gan et al., 2007). Monoethic methods divide the data set into clusters on the basis of a single pre-defined variable, whereas polyethic methods divide data based on the values of all variables (Gan et al., 2007). An example for a monoethic divisive algorithm is DIANA (DIvisive ANAlysis). This algorithm splits the biggest cluster at each step. The definition of the largest cluster can be, for example, on the basis of its diameter. The splitting process continues until each object is in a single cluster (Gan et al., 2007).

Both, agglomerative and divisive techniques have certain drawbacks. First, at an early stage incorrect grouped data points cannot be reallocated once the algorithm has continued to a

higher stage. Second, different similarity measures between two clusters might lead to different results (Gan et al., 2007). There are several methods for the representation of hierarchical clustering, which make human interpretation a lot easier. N-tree structures and dendograms are among the most common ones. The n-tree structure is a hierarchically nested tree diagram (see Figure 4.7). Terminal nodes or leaves are displayed with an open circle which represents a single data point. The internal nodes depicted by a filled circle represent a group or cluster (Gan et al., 2007). A dendogram is a valued tree (Gan et al., 2007) (see Figure 4.8). Each internal node is associated with a certain height which is based the clustering structure.

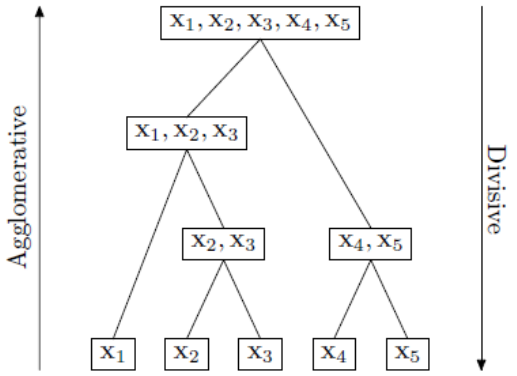


Figure 4.6: Divisive and agglomerative techniques in hierarchical clustering algorithms (taken from Gan et al., 2007)

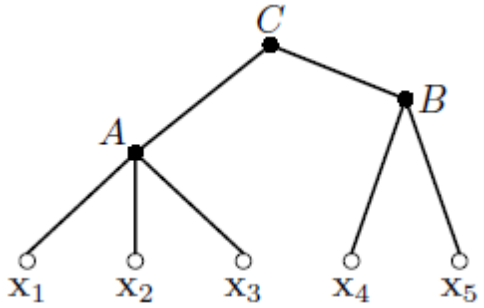


Figure 4.7: The n-tree, a representation in hierarchical clustering (taken from Gan et al., 2007)

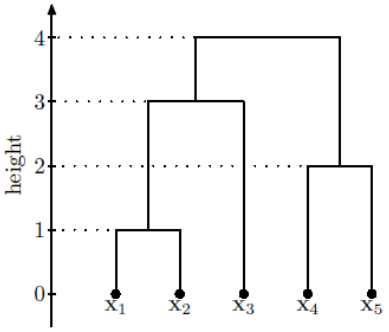


Figure 4.8: Dendogram (taken from Gan et al., 2007)

4.3.3.2 Partitional clustering

In partitional clustering a set of data points are assigned into k clusters without any hierarchical structure (Xu & Wunsch, 2009). The idea of cluster analysis is to summarize a partition of data in which data objects in the same clusters are homogenous, and data objects in different groups are distinct. This homogeneity and separation are evaluated with the criterion function. An example of a criterion function is the sum-of-squared-error criterion (Xu & Wunsch, 2009). The partition of data that minimizes the sum-of-squared-error criterion is optimal and is called the minimum variance partition (Xu & Wunsch, 2009). The sum of squares clustering model (SSCM) describes this process. The main objective of SSCM is to minimize the total weighted squared difference in cluster group membership. Thus, it tries to identify the optimal partition of entities. This concept is realized in the K-

means algorithm, which is one of the most popular clustering algorithms (Xu & Wunsch, 2009). The K-means algorithm can be applied to large-scale data sets especially when the resulting clusters are likely to be compact and hyperspherical (Xu & Wunsch, 2009). K-means assumes that the number of clusters K is already known by the user in advance, which is not true in practice. Thus, there are several heuristics, which can identify the value of K in advance.

4.3.4 Spatial and space-time cluster analysis in epidemiology

Cluster analysis is an important function of the data mining process. It has applications in many domains such as epidemiology. In epidemiology the distribution pattern of a disease depends on a few factors. These are environmental conditions, the population at risk, and the type of disease. Environmental conditions are important in terms of providing a germ an appropriate place to flourish. Thus, the success of arthropod-borne diseases such as malaria or the WNV depends on the prevalence of vectors and hosts. Thus, this kind of diseases can only exist in regions, providing space and habitats for arthropods and the hosts. In addition, it is necessary to know the situation of the population in an area where a disease is likely to emerge. Are there large settlements, such as towns, villages, or are there only a few scattered houses across a large area? The last factor, the type of disease, is important in terms of contagiousness and virulence. The distribution pattern of an influenza disease will differ significantly from the distribution pattern of a non-contagious encephalitis or of various cancer types.

The detection of clusters in a data environment deals with revealing those regions were some quantity is significantly higher than expected (Neill et al., 2005). The primary objective of spatial cluster analysis is to pinpoint the location, shape, and the size of a cluster (Neill et al., 2005). Additionally, it is necessary to determine whether a potential cluster is likely to be a true cluster. Hypothesis testing is applied in order to make sure that a potential cluster is not only a probable cluster, but a true cluster. Most cluster analyses are of purely spatial nature. However, in epidemiology the temporal dimension plays a crucial role in the detection of both retrospective and prospective clusters. Prospective clusters are clusters that have emerged at present and are still active, while retrospective methods include clusters which emerged and have ceased to exist in the past (Neill et al., 2005). According to Neill et al. (Neill et al., 2005) there are two methods of how a temporal component can be integrated into cluster analyses. The first approach is to carry out a purely spatial cluster analysis at each time step. This method, however, fails to detect more slowly emerging clusters. The second approach is to treat time as another dimension in the analysis. The drawback in this method is that it is likely to detect less relevant clusters. For this purpose, Neill at al. (2005) investigate the difference between the temporal and the spatial dimension. They reveal that unlike space, time has a point of reference, namely the present. But, temporal information lacks the so-called baseline denominator data, which might be the population are employed people. Instead, in temporal analysis there are expected values which can be derived from time series of past counts. Finally, time has an explicit direction, “proceeding from the past, through the present, to the future” (Neill et al., 2005). The general motivation of cluster analysis in epidemiology is to detect clusters which emerge over time.

4.3.5 Identification of spatial and space-time clusters in areal aggregated data

According to Anselin, a crucial aspect of pattern recognition, such as hot spot analysis is the determination of patterns on the map which reflect true clusters or outliers (Anselin et al., 2000). The counterparts of true clusters are “spurious” clusters, which are interpreted visually as clusters but in reality, they are no clusters. This is because the human mind always tries to identify patterns in clusters, even if the data is distributed perfectly randomly (Anselin et al., 2000). For this purpose, cartographic principles have been introduced, in order to ensure a proper interpretation.

Since hot spots share the features of both a geographic boundary and events within that boundary, Anselin states that the easiest way to identify hot spots is to partition an area into a fixed set of boundaries like for example zip-codes, and to develop a set of rules (Anselin et al., 2000). The “rule base” includes time intervals, threshold crime counts, and changes in crime counts. An example for a rule could be: If the crime counts in one zip-code exceed a certain threshold value, then it might be considered to be a hot spot. The boundaries can be chosen fixed or ad hoc. The disadvantage with fixed boundaries is that they are not very dynamical, but hot spots are. They may cross the fixed boundaries or vary in size (Anselin et al., 2000). An example for ad hoc clustering in point data is the nearest neighbor clustering in CrimeStat (Levine, 2004). The program provides two ways of visualizing the clusters. One way is to let the program draw ellipses around the cluster. The alternative is to let the program draw convex hulls by connecting all edge points of a cluster.

Hot spots are, by definition, small in area (Anselin et al., 2000). Thus, for analysis purposes large scales are used. However, when areal aggregated data are used their might be variations in the estimated effects of models arising from differences in the areal units (Anselin et al., 2000). This is known as the modifiable areal unit problem (MAUP). Thus, widely varying parameter estimates result from re-aggregating data by areal units of different sizes (Anselin et al., 2000).

Anselin (Anselin et al., 2000) defines exploratory spatial data analysis (ESDA) as a collection of techniques to describe and visualize spatial distributions. ESDA is about identifying spatial outliers, discovering patterns, clusters, or hot spots. Spatial autocorrelation is one principle of ESDA. It can be applied to both point data and areal data. In this thesis spatial autocorrelation is used to identify clusters in areal aggregated WNV data. The implemented methods (Moran Scatter Plot, LISA) determine the degree of deviation from spatial randomness in a study area.

4.3.6 Hot Spot (Cluster) Analysis Types

Hot spots or hot spot areas are concentrations of incidents within a limited geographical area that appear over time (Levine, 2005). In crime analysis the concept of hot spots is very useful, since police officers can focus their attention on particular environments where crime incidents peak. Levine (Levine, 2005) defines the concept of hot spots perceptual and not existing in reality. Nevertheless, there could be areas where there is sufficient concentration of certain activities which are labeled as areas of high concentration. Hot spot analysis is statistically known as cluster analysis. Several techniques have been developed on how to detect and analyze hot spots (Levine, 2005). The most intuitive type of cluster is, when only the location of incidents is considered. Thus, the location with the highest number of incidents is considered to be a hot spot. Hierarchical techniques were already pointed out in

chapter 4.3.3.1. They act like an inverted tree diagram and summarize features on the basis of a specific criterion, such as nearest neighborhood. Thus, the technique creates groupings or clusters of first order, second order, and so on. Partitioning techniques group incidents in a pre-specified number of groupings (Levine, 2005). The most popular partitioning technique is the K-means technique. Another routine to identify hot spots is the density technique. This technique searches for dense concentrations of incidents. The risk-based technique identifies clusters in relation to an underlying base at risk, such as population, or employment (Levine, 2005). Apart from all these techniques are clumping and miscellaneous techniques. Most of the hot spot techniques require event or incident data as input information. Unlike the Nearest Neighbor Hierarchical Clustering or the K-means algorithm, the Local Moran's I technique, developed by Anselin, however, requires data aggregated by zones (Levine, 2005). Apart from hot spots (a neighborhood with high intensity values), the Local Moran's I also provides information about cold spots (a neighborhood with low intensity values).

4.3.7 The self organizing map (SOM)

A self organizing map is an architecture or algorithm for an artificial neural network (Kohonen, 1990). According to Kohonen it is capable of creating spatially organized "internal representations" of various features. The SOM is particularly successful in pattern recognition tasks. Thus, it has found its application in various fields, such as robotics, process control, telecommunications, and speech recognition. Apart from feedforward and feedback networks, the self organizing map belongs to the third category in the field of network architectures for modeling nervous systems (Kohonen, 1990). In this category learning is called competitive, unsupervised, or self organizing. The principle goal of the SOM is to transform an incoming signal pattern into a one or two dimensional discrete map (Guo et al., 2006). The self-organization is created by neighboring cells, which compete in their activities. Subsequently, they develop adaptively into specific detectors of different signal patterns. The cells become tuned to various input signal patterns. Following those signal patterns, the locations of the cells become ordered. Thus, the spatial location of a cell in a network, then, corresponds to a particular domain of input signal patterns (Kohonen, 1990). Eventually, a coordinate system for the input features is created. The SOM forms the required topographic map of the input patterns (Guo et al., 2006).

Guo et al. (Guo et al., 2005) adapt the SOM for multivariate analysis and geo-visualization. A challenge in those fields is the high-dimensionality of data. The SOM is capable of projecting high-dimensional data to a low-dimensional space while still preserving nonlinear relationships. Thus, SOMs are a method of abstraction or summarization because of their ability to compress information. For this purpose, SOMs conduct a many-to-one projection, so that more than one data item in the input data can be projected to the same node if they are similar (Guo et al., 2005). In geographic analysis SOMs have a wide range of application. Guo et al. (2005) mention a few. Those include: visualization of patterns in census data, spatialization of non-spatial information, and exploration of health survey data.

Kohonen (Kohonen, 1990) suggests that so-called brain maps have a lot in common with SOMs. Thus, he concludes that the internal representation of information in the brain is organized spatially. Modern imaging techniques which use radioactive tracers have revealed a fairly detailed organizational view of the brain. In higher animals, "various cortices in the cell mass contain many kinds of map" (Kohonen, 1990). So, for example, in the visual areas

there are line orientation and color maps. In the auditory cortex there are “tonotopic maps”. There are even parts of the brain which visualize a representation of the face and the body in form of a map (Kohonen, 1990).

4.4 Selected techniques for cluster analysis

Clusters and a clustering effect in the study areas both in the U.S. and in the state of Louisiana can be detected with the above listed software packages. The packages are quite different as far as their structure and functionalities are concerned. However, each software package offers some specific algorithms in order to identify clusters in a data set. Some essential techniques which are part of the software packages used for analysis in this thesis are global and local spatial autocorrelation, retrospective and prospective tests, and the scan statistic.

4.4.1 Spatial autocorrelation

Tobler’s first law of geography says that: “Everything is related to everything, but near things are more related than distant things” (Tobler, 1970). This effect is called positive spatial autocorrelation. Waller and Jacquez talk about spatial autocorrelation when near rates tend to be similar (Waller & Jacquez, 1995). In their description of disease models they refer to either spatial autocorrelation or spatial heterogeneity, when mean regional rates vary from place to place (Waller & Jacquez, 1995). This effect is also referred to as negative spatial autocorrelation. Waller and Jacquez (1995), furthermore, explain spatial heterogeneity in epidemiology as contrasting independent disease cases arise from a noninfectious disease. In addition, the disease rate varies across the study area. Under this point of view, spatial autocorrelation would occur among infectious disease cases, where the rate is constant across the study area (Waller & Jacquez, 1995). Apart from positive and spatial autocorrelation there is complete spatial randomness (CSR) or no spatial autocorrelation. Diggle’s hypothesis of CSR is that there are no interactions or relations amongst the events and the intensity of events does not vary over the study area (Diggle, 2003). Griffith (Griffith, 2009) gives a number of real-world examples for spatial autocorrelation. For example, minerals cluster at certain locations in the Earth’s surface and are not ubiquitous. When talking about the real estate market house value assessments are established on the basis of similar nearby houses. In epidemiology, a disease like the WNV creeps across an area through arthropod-borne contagion. Griffith describes the diffusion of the WNV over the U.S. as a “spatial process mechanism”. Like a weather front, it emerged in Long Island in 1999 and quickly diffused westward throughout the remainder of the country. A weather front can result in highly spatially auto-correlated local weather conditions (Griffith, 2009). This was similar to the spread of the WNV across the U.S. Spatial autocorrelation cases can be divided into different relationship tendencies between adjacent values on a map. Exceptionally, remotely sensed images almost always display a very strong positive spatial autocorrelation (Griffith, 2009). Moderate positive spatial autocorrelation can be discovered in population maps. This effect is caused by humans’ tendency to form groups to live in and establish settlements with a central administration unit. Thus, most socioeconomic or demographic data display a moderate positive relationship (Griffith, 2009). When it comes to negative spatial autocorrelation, Griffith talks about a “geographic competition” among values in the study area. He creates an index of local competition of European Union members in terms of land sizes. For this purpose, he creates a ratio of each country’s actual size and the corresponding Thiessen polygon. The ratio reveals a negative spatial

autocorrelation. Countries like Luxembourg, Slovenia and the Czech Republic of very low ratio values are surrounded by countries with very high ratio values (Griffith, 2009).

The concept of spatial autocorrelation has been incorporated into two commonly used models: The Moran Coefficient (Moran's I) and the Geary Ratio (Geary's C). In the analyses conducted in this thesis only the Moran Coefficient (MC) will be used and explained in detail.

4.4.1.1 Global spatial autocorrelation

The Moran Scatterplot displays the global relation trend in a study area. It can be divided into four quadrants. The values in the study area are assigned to one of the four quadrants depending on their rate and the rates of adjacent objects (polygons) (see Table 4.3). Each quadrant corresponds to a specific type of autocorrelation. In the course of the calculation of a Moran Scatterplot a z-transformation is conducted in order to make values comparable to other values.

Quadrant Number	Abbreviation	Explanation
I	H-H	Investigated polygon has a high rate as well as its neighboring polygons. The occurrence of Hot Spots is likely.
II	H-L	Investigated polygon has a high rate but the neighboring polygons have low values. The occurrence of spatial outliers is likely.
III	L-L	Investigated polygon has a low rate as well as its neighboring polygons. The occurrence of Cold Spots is likely.
IV	L-H	Investigated polygon has a low rate but the neighboring polygons have high values. The occurrence of spatial outliers is likely.
H-H…High-High, H-L…High-Low, L-L…Low-Low, L-H…Low-High		

Table 4.2 Quadrants of the Moran Scatterplot

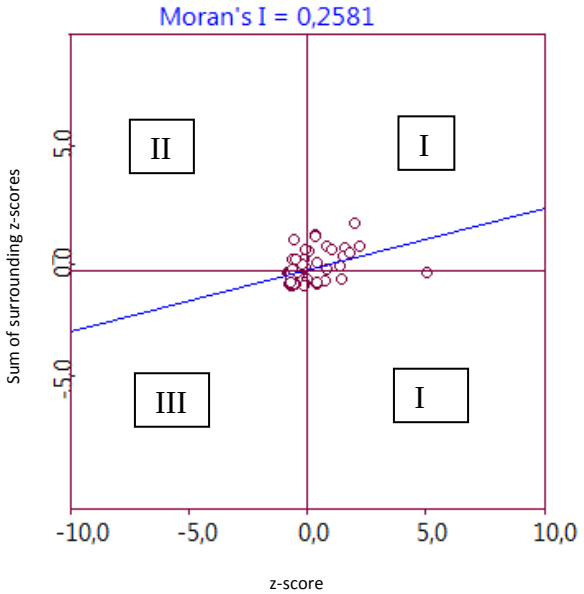


Figure 4.9: Moran Scatter plot for WNV incidences in Louisiana in 2002

Figure 4.9 displays a Moran Scatter plot to observe spatial autocorrelation among WNV incidences in Louisiana in 2002. The corresponding characteristics of each quadrant are listed in Table 4.3. Most of the values appear in quadrant I or III. Thus, there might be hot and cold spots, which indicate that values with similar rates are close to the central polygon. This characterizes of positive spatial autocorrelation. Moran's I is 0.2581, resulting in a slightly positive spatial autocorrelation.

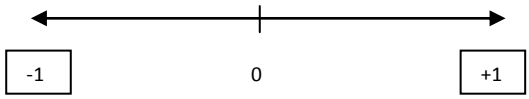


Figure 4.10: Interpretation of Moran's I

The Moran's I statistic can be a number between -1 and +1 displayed in Figure 4.10. The value +1 is for maximum positive spatial autocorrelation and the value -1 stands for maximum negative spatial autocorrelation. There is no spatial autocorrelation if Moran's I is a slightly negative value very close to zero.

4.4.1.2 Local Moran's I spatial autocorrelation (LISA)

The basic concept of Anselin's Local Moran's I is that of a local indicator of spatial association (LISA) (Levine, 2005). The LISA is able to indicate for each observation the extent to which there is significant spatial clustering of similar values around that observation (Levine, 2005). LISA is a value, which determines the similarity between one observation and its neighbors. In order to detect hot spots LISA tests whether an observation in a neighborhood with high intensity values is similar (high) or distinctly different (low).

The Moran scatter plot can be visualized with a LISA map containing neighborhood information. Local spatial autocorrelation yields a measure of spatial autocorrelation for each individual location (Anselin, 2003b). In a thematic map the regions with a significant Local Moran's I statistic are highlighted. Depending on the statistic each significant region is assigned to a category, deriving from the Moran scatter plot (see Table 4.3). The remaining regions belong to the category "not significant". The essential step for the creation of LISA maps is a weights file on the basis of either contiguity or distance (Anselin, 2003b). Distance weights files require exact coordinates of each item in a data set. This can be either x and y coordinates when working with a point pattern or centroids for polygon data. The number of neighbors of each polygon depends on the determined threshold distance. Each data item should at least have one neighbor, thus, the threshold distance should be set corresponding to that criterion.

As far as contiguity weights are concerned, the specification of neighbor relationships can influence drastically the outcome of statistical analyses (BioMedware, 2012a). There are two possibilities to determine the neighborhood of polygons: rook and queen (see Figure 4.11). These two options are named after movements on a chessboard (BioMedware, 2012a). Queen contiguity results in significantly more neighbors, since a possible candidate is considered as neighbor when only sharing one vertex. In contrast, when choosing rook contiguity the polygons have to share an edge to be considered as neighbors.

Rook Contiguity Queen Contiguity

Figure 4.11: Two different options of contiguity (taken from BioMedware, 2012)

A common way to represent neighbor relationships is the spatial weights matrix (BioMedware, 2012b). It is an nxn matrix, where n is the number of data items in the study area. Open GeoDa generates the matrix in form of a weights file, listing up the central polygon's ID, the number of neighbors, and the ID's appertaining of those neighbors (see Figure 4.12).

```
0 64 La OBJECTID
1 6
59 52 36 32 8 23
2 3
54 27 6
3 5
51 42 10 7 40
4 3
60 33 44
5 5
61 38 23 26 43
```

Figure 4.12: Excerpt of the weights file created by Open GeoDa

4.4.2 The Global and the Local score statistic

Global statistics are conducted in order to get a summary of the entire study region. General tests are carried out with global statistics. The result is a single summary statistic, which characterizes any deviation from the null hypothesis of a random pattern (Rogerson, 2005). By contrast, local statistics are carried out for subsets of the study area. Thus, many local tests can run simultaneously in order to detect clustering when there is no idea of the location of possible clusters in advance. In common, these tests are referred to as tests for the detection of clustering (Rogerson, 2005). A second possibility is to carry out a local statistic to find out whether clustering occurs around particular foci. These are called focused tests (Rogerson, 2005). Anselin argues that local statistics have the property to sum up the respective global statistic (Rogerson, 2005). Considering that, if a general test does not find a significant deviation from spatial randomness it is still useful to carry out local statistics. The local statistics may still reveal isolated hotspots of increased incidence, which are, however, insufficient to lead to a global signal (Rogerson, 2005). When the global statistic reveals significance in clustering, local statistics can provide additional information. Thus, local statistics help to decide whether the study area is homogeneous in terms of a similarity of local statistics throughout the area or whether there are a few very strong local outliers (Rogerson, 2005). If local statistics are similar across the study area they contribute

approximately equally to the significance of the global signal, whereas in the second case a few primary outliers lead to the significant global statistic. Local statistics, therefore, have the power to reveal details of the data distribution on a more detailed level. Fuchs and Kenett (Fuchs & Kenett, 1980) derived a test based on the largest outlier in a distribution, the M test. The M test has the power to determine those data items which are responsible for the rejection of the null hypothesis of spatial randomness. The M test is especially useful in situations when there are a small number of outlying items in a data set (Rogerson, 2005).

Local Score Statistic U_i (focused)	Global Score Statistic U^2
$U_i = \sum_{j=1}^{m} w_{ij}(O_j - E_j)$	$U^2 = \sum_{i=1}^{m} U_i^2$

Table 4.3: Formulas for the local and the global score statistic (taken from Rogerson Rogerson, 2005)

The focused local score statistic (U) depicted on the left in Table 4.4 tests for raised incidence around a region i. The U statistic has the power to reject false null hypotheses. Under the null hypothesis this statistic has an asymptotic normal distribution with mean zero (Rogerson, 2005). The weights w_{ij} are chosen as a function of the distance between the regions i and j. The variable m stands for a set of regions. O_j are the observed cases in a region and E_j are the expected cases in that region. In the right hand side of the table the global statistic described by Rogerson (Rogerson, 2005) is highlighted. It is based on the local score statistic. Again, m stands for a set of regions.

If the expected and observed cases for each sub region are available they can be transformed into a standard normal variable via a variety of transformation techniques. Three transformations are adopted in GeoSurveillance (see chapter 4.2.2) for retrospective testing: Poisson-based, Freeman-Tukey and Rossi transformation. The Poisson based transformation is equal to the local score statistic in Table 4.4.

4.4.3 Cumulated Sum (Cusum) Control Charts Methods

Originally developed for quality control of industrial processes, in GeoSurveillance the method was adapted for prospective tests of spatial cluster detection. The cusum methods repeatedly update associated test statistics as new data becomes available (Rogerson et al., 2009). This is useful in order to record deviations of the mean of a variable of interest. When talking about deviations, these can be both decreases as well as increases in a variable value. In epidemiology, the focus is set on value increases. Also the cusum method realized in GeoSurveillance is specifically sensitive to variable increases. The cusum method is applied to individual sub regions simultaneously to detect any increase in regional observations as quickly as possible.

The cumulated sum is the sum of the differences between the values and the mean (Taylor Enterprises, 2012). When the values are above average during a certain period of time, the amounts added to the cusum will be positive and the sum increases (Taylor Enterprises, 2012). On the other hand, values below the mean will decrease the sum. Rogerson, Yamada and Lee define the cusum method in GeoSurveillance as follows:

$$S_{i,t} = \max\left(S_{i,t-1} + z_{i,t} - k, \quad 0\right),$$
$$S_{i,0} = 0,$$

Formula 4.2: Cusum formula (taken from Rogerson et al., 2009)

Since the mean is subtracted from each value, the cusum also ends at zero ($S_{i,0} = 0$). $O_{i,t}$ are the approximately normally distributed values in a region i at a certain time period t. These are standardized to $z_{i,t}$ using the mean equal to the expected number of cases $E_{i,t}$ and a known variance. The parameter k is usually set to 0.5 in order to minimize the time to detect an increase of 2 k standard deviation in the mean of $z_{i,t}$. When the cusum exceeds the given threshold parameter h a significant increase in the subregion i is signaled (Rogerson et al., 2009).

4.4.4 Kulldorff's Scan Statistic

A scan statistic is used to detect clusters in a point process (Kulldorff, 1997). If only areal aggregated data are available, the coordinates of centroids or important locations in the respective unit, like for example county seats, can be used for calculation. While most statistical methods for cluster analysis of a spatial point process can either detect the location of clusters or do inference about it, the scan statistic can do both. Thus, the scan statistic answers the question whether there are clusters in a data set or not and it also reveals the geographic location of the cluster. At this point it detects both clusters with exceptional high values as well as clusters with conspicuous low values. The scan statistic, however is not capable of answering the question whether there is a clustering effect over the study region as a whole, like for example what might be the case with a spreading infectious disease (Kulldorff, 1997).

The underlying technique of the scan statistic is a predefined measure, which can be either a circular or an elliptic window that keeps scanning the entire study area in the search for clusters. If one aims to also integrate a temporal component into the cluster detection, then the measure has to be extended in a third temporal dimension. What has to be taken in consideration when integrating time are changes which might happen across time, like for example changes in the population. When looking for space-time clusters instead of the circular window a cylinder is used to incorporate the third, the temporal dimension, as well. The maximum number of points in the window is recorded and compared to its distribution under the null hypothesis of a purely random Poisson process (Kulldorff, 1997).The Poisson process is used in order to predict the degree of spread around a known average. The Scan Statistic implies two models: Bernoulli and Poisson. It depends on the input data which model would be most appropriate for cluster detection. The Bernoulli model is designated for binary counts. An example for the application of the Bernoulli model is the investigation of the Sudden Infant Death Syndrome. There is the birth rate per state and the disease rate, which form both a binary count. The Poisson model, by contrast, finds its application with counts relating to some continuous factor (Kulldorff, 1997). Both models determine an exactly zone Z which is a subset of the entire study area G. Each individual has a probability p of being a point in the zone Z or has the probability q of being located outside the zone Z. Considering, that, the null hypothesis H_0 and the alternative hypothesis H_A can be stated as follows (Kulldorff, 1997):

H_0: p = q

$$H_A : p > q$$

The null hypothesis of no spatial clustering is rejected, when the scanning window identifies a cluster. Apart from the most likely cluster the scan statistic is also able to reveal secondary clusters with high likelihood values. In this context, most secondary clusters will be overlapping with the most likely cluster. More interesting are secondary clusters which emerge in a different region of the study area.

4.4.5 Visualization of multivariate clusters over space and time

The visualization of multiple perspective data requires methods that can simultaneously visualize spatial, temporal and multivariate patterns. In crime analysis, for example, it is important to include attributive data like socioeconomic factors or crime types apart from the space-time components. Guo et al. (Guo et al., 2005) describe several challenges when it comes to multivariate spatial data analysis. One of them is that the high dimensionality of geographic data sets can cause serious problems for analysis methods. A dataset with high dimensionality is a data set with a large number of variables, whereas a data set with a large number of cases it is referred to as a large dataset (Guo et al., 2005). Also, large and high-dimensional data sets demand that analysis methods are computationally efficient. Another aspect is that it is unlikely for all attributes to interrelate meaningfully. Additionally, to interpret the meaning of patterns expert knowledge is required. In other words, both visualization as well as computational methods, individually used, are insufficient in terms of analysis of high-dimensional data.

Furthermore, Guo et al. (Guo et al., 2005) describe three major factors why the detection of patterns in data can be hard to discover. The first difficulty arises due to high dimensionality in the data. One premise for analysis is that all variables in the input data are meaningful and relevant to each other. A data set might also include several different patterns, created by a subset of variables respectively. Thus, it is important to initially know the items of the subset in order to make sure that they are meaningful to each other. The second aspect which has to be taken in consideration in analyses is that potential patterns may take various forms. The third factor is rather a challenge than a factor. It is about the visualization of multivariate geographic patterns. Maps are an essential component in the visualization for geographic patterns. What makes pattern detection more sophisticated is to combine the detection of geographic patterns with multivariate pattern detection.

Guo et al. (Guo et al., 2005) propose an integrated geographic knowledge discovery environment. This system is able to visualize patterns in both the geographic space and the attributive space. In addition, it supports human interaction and interpretation, facilitating the examination and explanation of patterns. The environment consists of several major components including data processing, unsupervised feature selection, multivariate analysis with the SOM, multidimensional visualization and multivariate geographic visualization in form of a map (Guo et al., 2005). In Guo's approach he uses dimension reduction methods in order to map and visualize patterns across multiple variables and dimensions. For this purpose, Guo developed a multivariate mapping approach, called SOMVIS. This approach includes a SOM, which itself is a dimension reduction and clustering method. Guo et al. (2006) extended the SOMVIS to accommodate a temporal dimension. The resulting software environment VIS-Stamp is able to perform multivariate space time pattern analysis.

5 Results

This section is divided into the results for the U.S and the results for Louisiana. First, the results for the U.S. software per software are introduced, then the results for Louisiana are displayed.

5.1 Spatial and Temporal analysis of the WNV in the United States

In this section the results of the analysis in the U.S. are displayed. First, the results of GeoDa are presented, then the results of SaTScan and Vis-Stamp.

5.1.1 Exploratory spatial data analysis (ESDA) in Open GeoDa

First, the spatial distribution of WNV incidents in the U.S. is investigated. Open GeoDa provides functionalities to determine the degree of spatial autocorrelation in the data. However, it lacks the functionality of integrating the temporal component as an additional dimension. Thus, the analyses are conducted for each year, starting with the first emergence of the WNV in New York in 1999. The last period of investigation is the year 2011. The connectivity chart reveals the neighborhood relations in the study area, including all U.S. states without Hawaii and Alaska plus the District of Columbia (see Figure 5.1). The chart categorizes U.S. states on the basis of their respective number of neighbors. The chart depicts that the majority of states have between four and six neighbors. The contiguity weights file was created with the Rook Contiguity option, thus, only those adjacent states which share at least an edge are considered to be neighbors. The LISA maps are created using Empirical Bayes (EB) Rates. This means that the input variable (total number of WNV cases) is normalized by the state population.

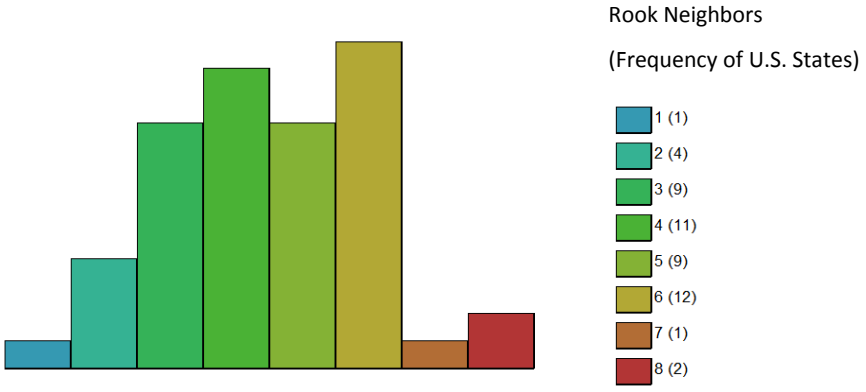
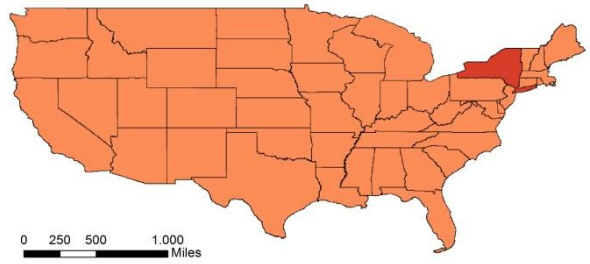
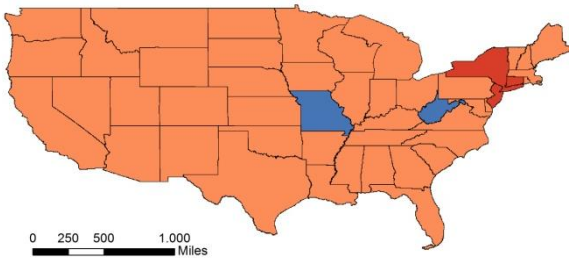
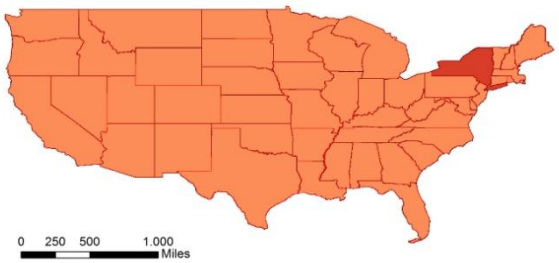
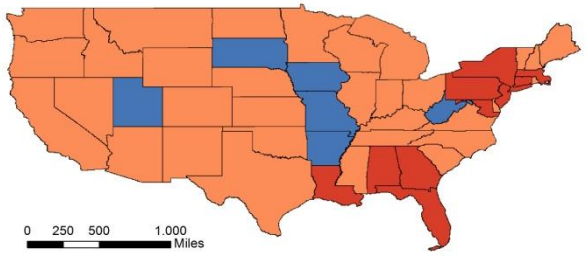
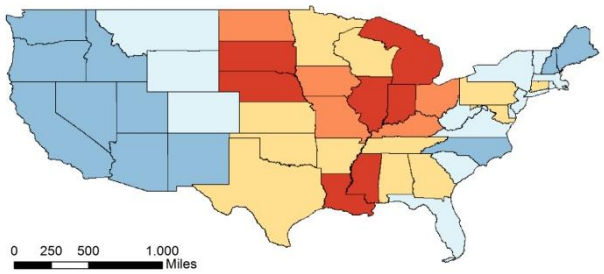
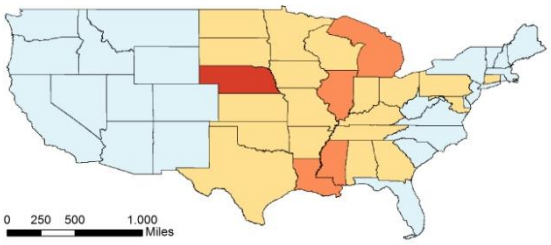
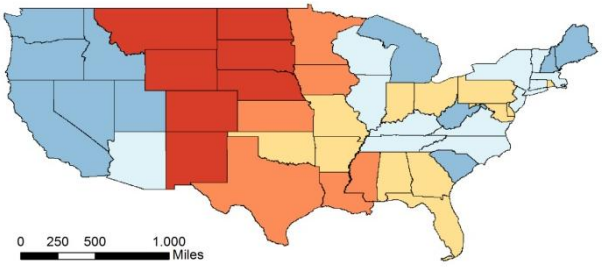
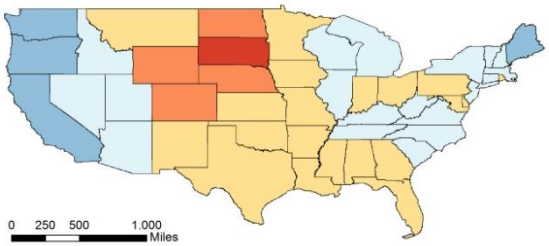


Figure 5.1: Connectivity chart of U.S. states

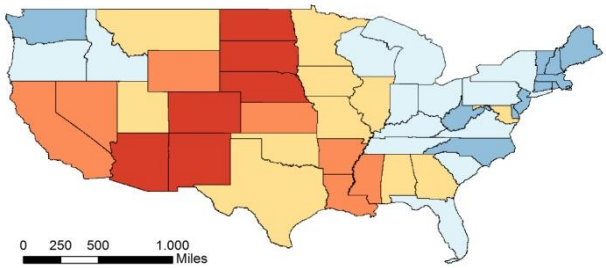
In order to detect statistical outliers in the study area outlier maps are created. Outlier maps highlight data values, which significantly deviate from the mean. Open GeoDa offers two kinds of outlier maps, namely the box plot map and the percentile map. The box plot map is the spatial equivalent of a box plot. Percentile maps are based on a simple data ranking and highlight extreme values which are located at the bottom and at the top of a data distribution. These techniques are implemented in order to get a better understanding of the data distributions. For the analysis, a hinge of 1.5 is used. The hinge criterion determines how extreme observations need to be that they are classified as outliers. Therefore, Open GeoDa provides two options: a hinge of 1.5 or a stricter hinge of 3. Open GeoDa provides five methods to produce choropleth maps for variables that are expressed as rates. In this analysis the total numbers of WNV cases are smoothed with the method of raw rates.

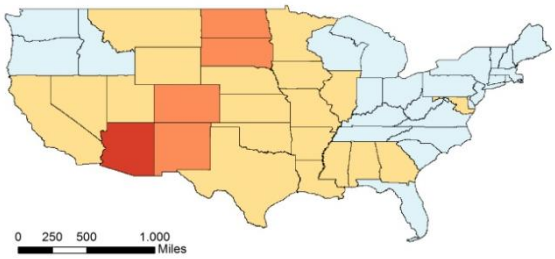
Box Plot Maps Smoothed with raw rates Hinge: 1.5	Percentile Maps Smoothed with raw rates
1999	
	
2000	
	
2001	
	
2002	
	
2003	



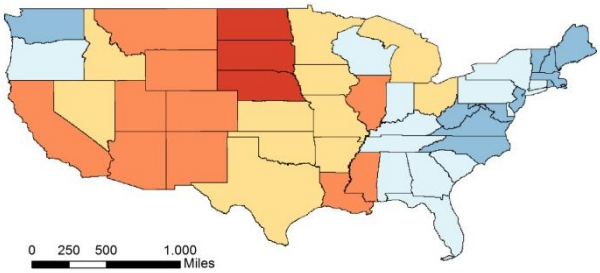


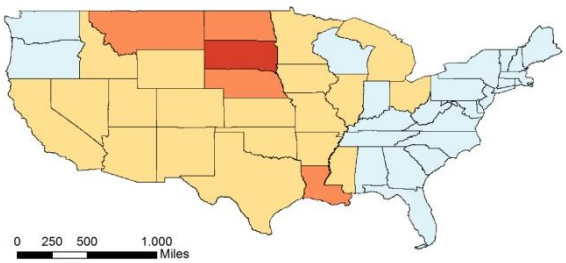
2004



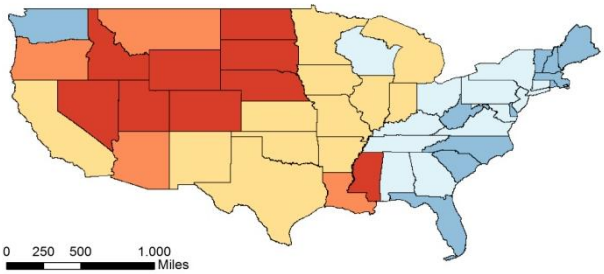


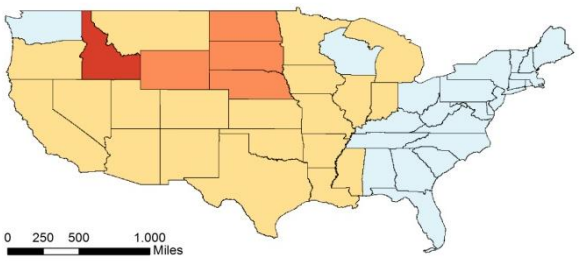
2005



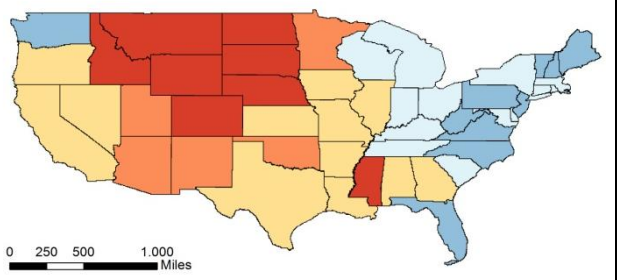


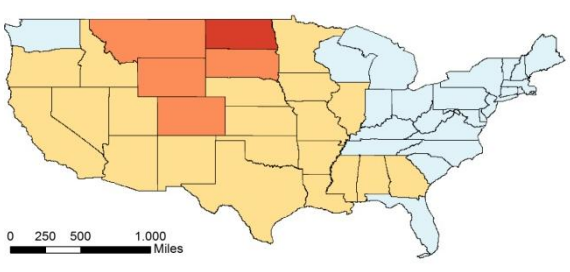
2006



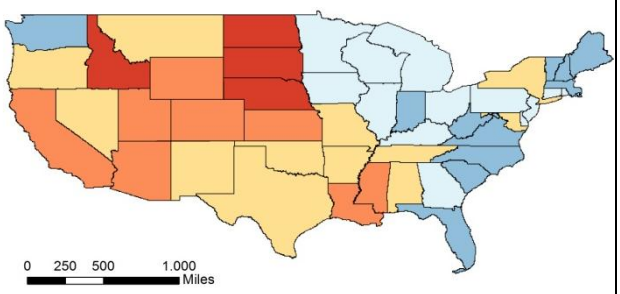


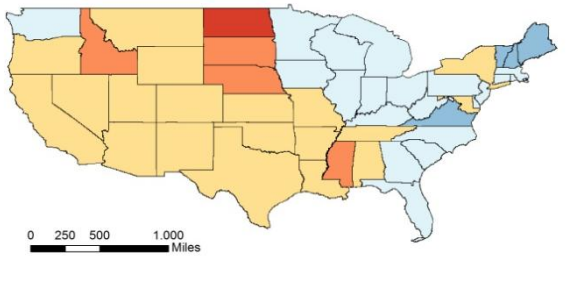
2007



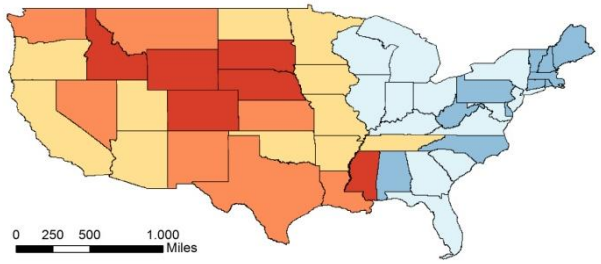


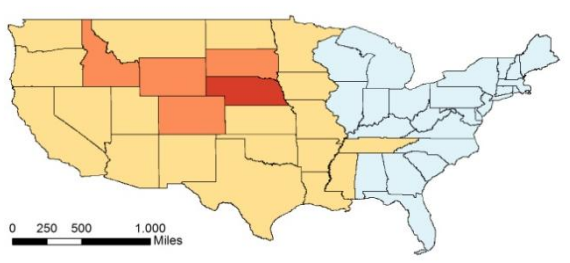
2008



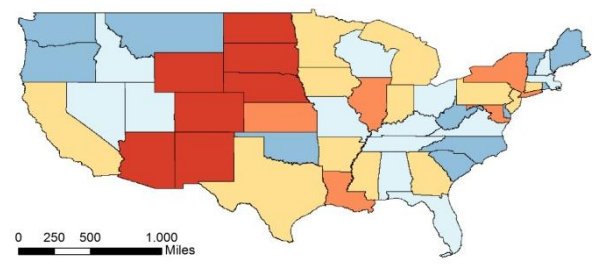


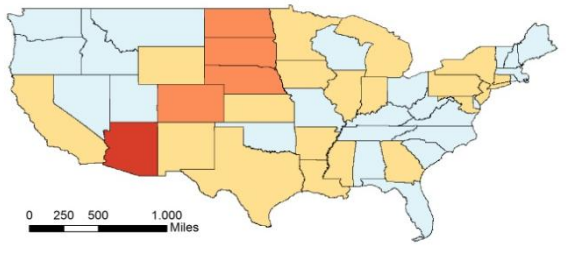
2009





2010





2011

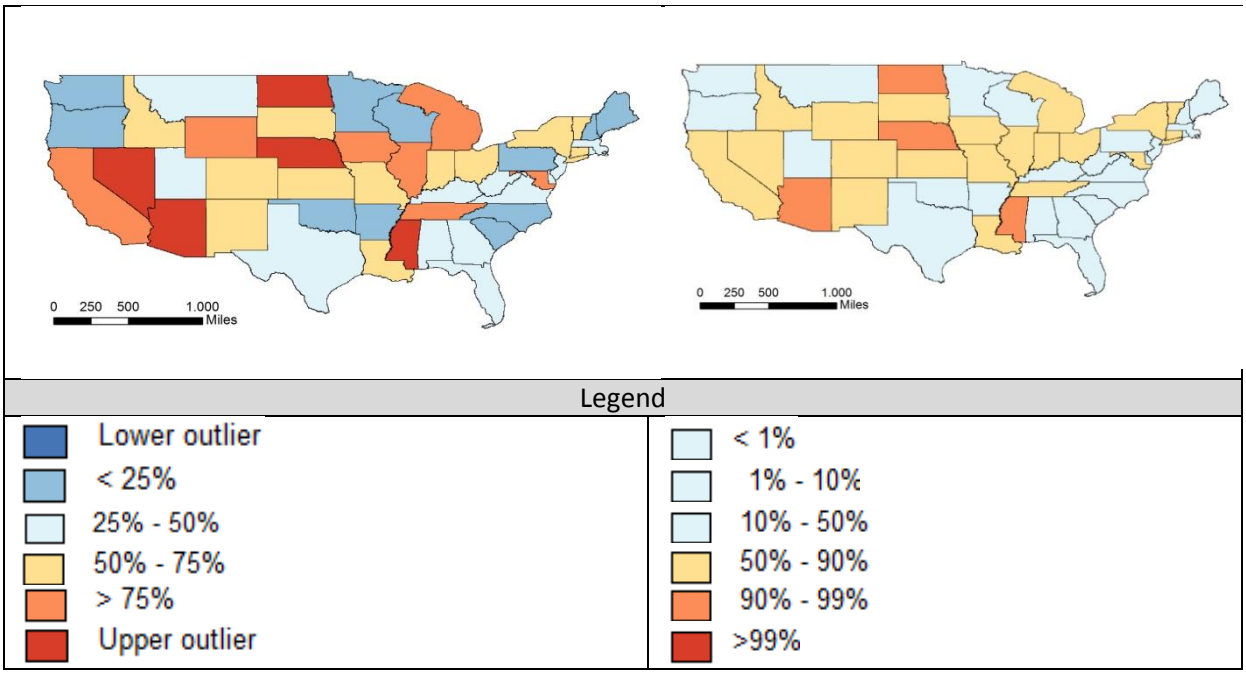


Table 5.1: Outlier maps of WNV disease cases in the U.S. in the study period from 1999 - 2011

The outlier, i.e., box plot and percentile maps reveal a similar pattern to the LISA maps. In the first two years of examination, Open GeoDa, incorrectly classifies the states with zero cases into the second highest categories. The 1999 and 2000 maps should thus be interpreted with caution. In the first three years (1999-2001), upper outliers are located on the north-eastern part of the U.S. Apart from 2002 upper outliers can be found in the central U.S. In 2003, 2006, 2007, and 2010 there are especially many upper outliers in the central and southern U.S. This gives an idea of how much the number of WNV cases actually differs from the mean. Though, this kind of visualization gives no information about the intensity (i.e. absolute numbers of disease cases) of the virus. In other words, in 2006 upper outliers are states with more than 500 cases, while in 2010 already states with more than one hundred cases are considered to be outliers. After the year 2001 there are no lower outliers any longer, but only states falling into the second lowest, i.e. <25% category. A cluster of such low numbers of WNV cases can be detected in 2002 on the west coast, when the virus has not yet reached the western part of the U.S. One interesting aspect is, that while the number of WNV cases becomes lower in most of the eastern states, two states in the south east are still having very high numbers of WNV cases, including Mississippi and Louisiana. The latter state is classified as upper outlier twice at the beginning of the study period. Mississippi is occasionally an upper outlier throughout the entire study period. The “success” of the virus in those two southern states might be a result of the hot, humid climate in the summer and the aspect that swamps act as an ideal breading habitat for both mosquitoes and avian species.

In 1999 the virus first entered the U.S., causing 62 disease cases, among those were 7 fatal cases in the state of New York (see Figure 5.2). In the LISA map the states Vermont, Massachusetts, and Connecticut are categorized as low-high spatial outliers due to the high number of cases in the “outbreak” state of New York and the low number of cases in the three neighboring states. The Moran Scatter plot in Figure 5.2 depicts a very slight negative spatial autocorrelation (Moran’s I = -0.0324). This value, however, is so close to zero, that it might be interpreted as no autocorrelation (i.e., spatial random distribution).

1999

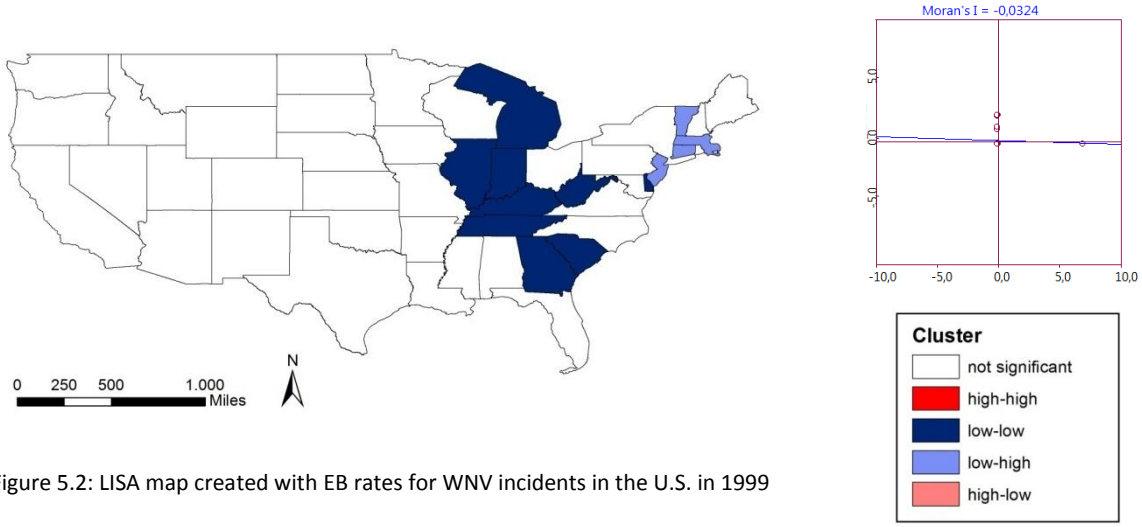


Figure 5.2: LISA map created with EB rates for WNV incidents in the U.S. in 1999

In the following year, in the state of New York there were significantly fewer reported disease cases than in the season before. However, the virus also entered the adjacent states New Jersey and Connecticut, leading to very few disease cases in those two states. Since there were no cases reported in the rest of the nation, the few disease cases in the northeastern part of the U.S. led to a high-high cluster. The Moran Scatter plot reveals a slight positive autocorrelation. Most of the states lie in the first and third quadrant. In this kind of distribution it is likely to have cold and hot spots. This is the case in the LISA map. There is a hot spot around New York State and a large cold spot to the west of that hot spot.

2000

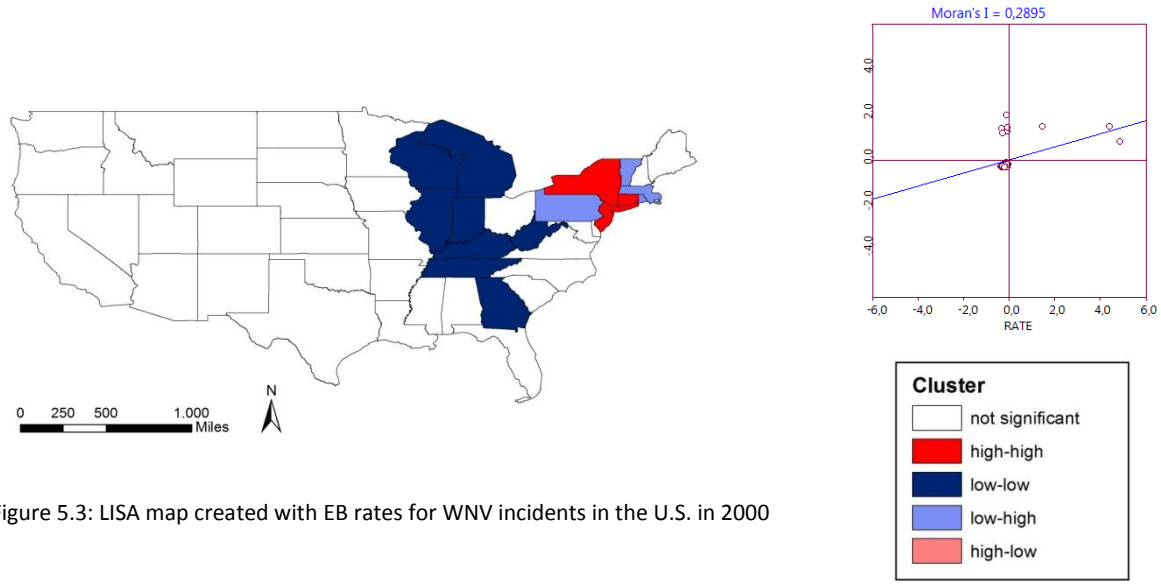


Figure 5.3: LISA map created with EB rates for WNV incidents in the U.S. in 2000

In 2001 the hot spot around New York State becomes larger, including now two additional states, namely, Massachusetts and Pennsylvania. New WNV cases emerge in Maryland. Apart from now the virus will spread continuously westward.

2001

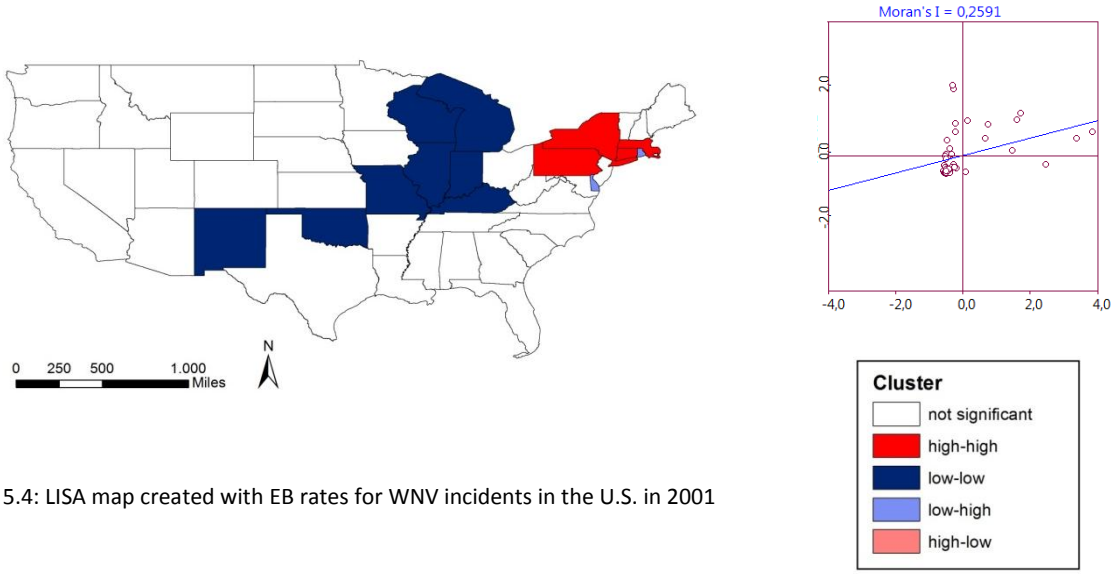


Figure 5.4: LISA map created with EB rates for WNV incidents in the U.S. in 2001

In 2002 the WNV has "conquered" large parts of the continental U.S. It was the most "successful" year for the virus since its emergence in 1999. Especially, neuroinvasive and non-neuroinvasive disease cases among humans exploded. In the state of Illinois more than 800 disease cases were reported to the CDC (CDC, 2011a), among those more than 500 neuroinvasive cases. Meanwhile, on the north-east coast things the number of disease cases was comparatively low, leading to a low-low cluster around the state of Connecticut. Still, the virus has not yet impacted the west coast, where a low-low cluster has evolved.

2002

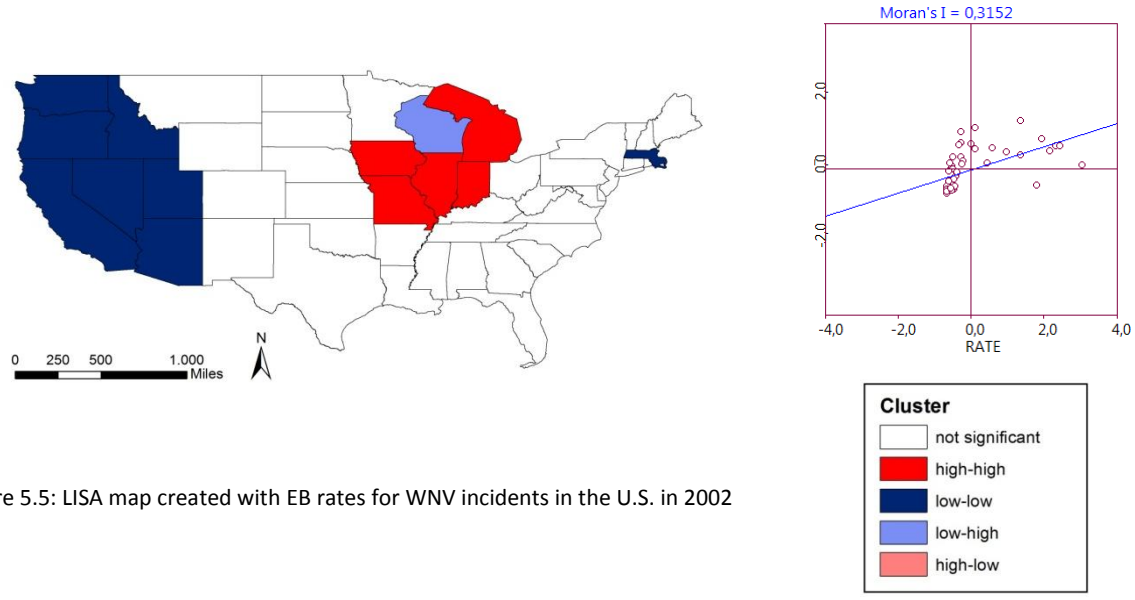


Figure 5.5: LISA map created with EB rates for WNV incidents in the U.S. in 2002

Regarding the Moran Scatter plot, the positive spatial autocorrelation has become more positive since 1999. That means that it is more likely to find clusters, either cold spots or hot spots in the LISA maps. In 2003 WNV incidents concentrate in the central and the northern U.S. There is a large hot spot, including the states of Montana, Wyoming, Colorado,

Nebraska, South Dakota, and North Dakota. In those states, as well as in a few surrounding states the virus reaches an unprecedented intensity in terms of contagiousness and aggressiveness.

2003

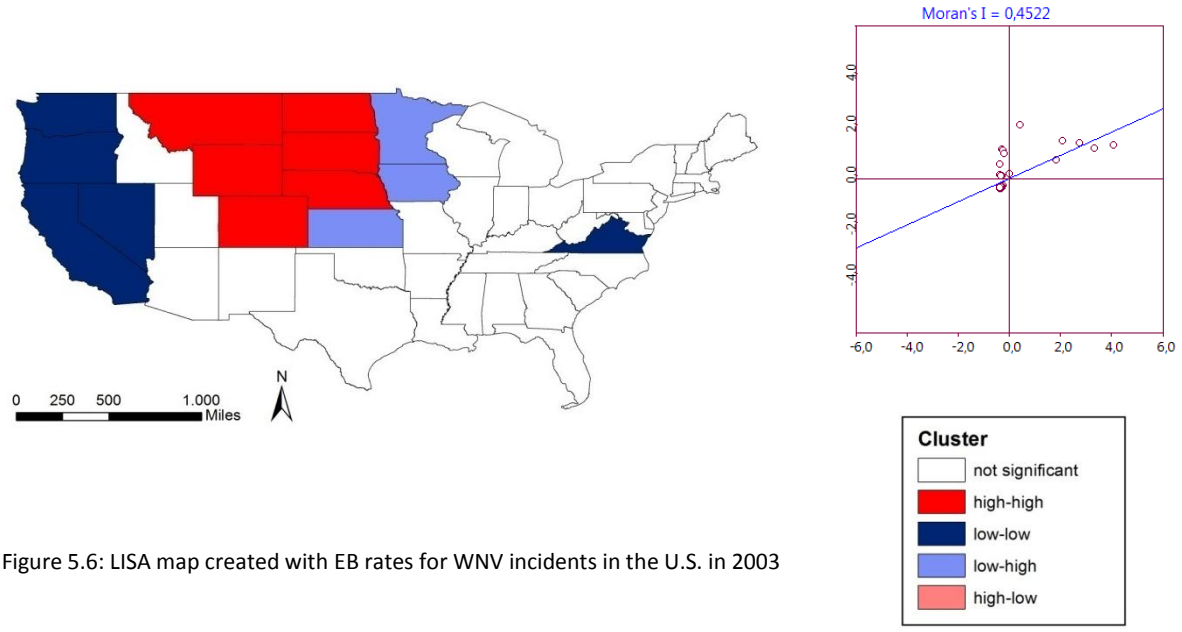


Figure 5.6: LISA map created with EB rates for WNV incidents in the U.S. in 2003

In 2004, the WNV has spread all across the U.S. However, the overall situation has become less severe. Thus, there are several hot spots, but they do not reach a high intensity as far as reported cases are concerned. California is an exception. More than 700 cases were reported throughout the state in 2004.

2004

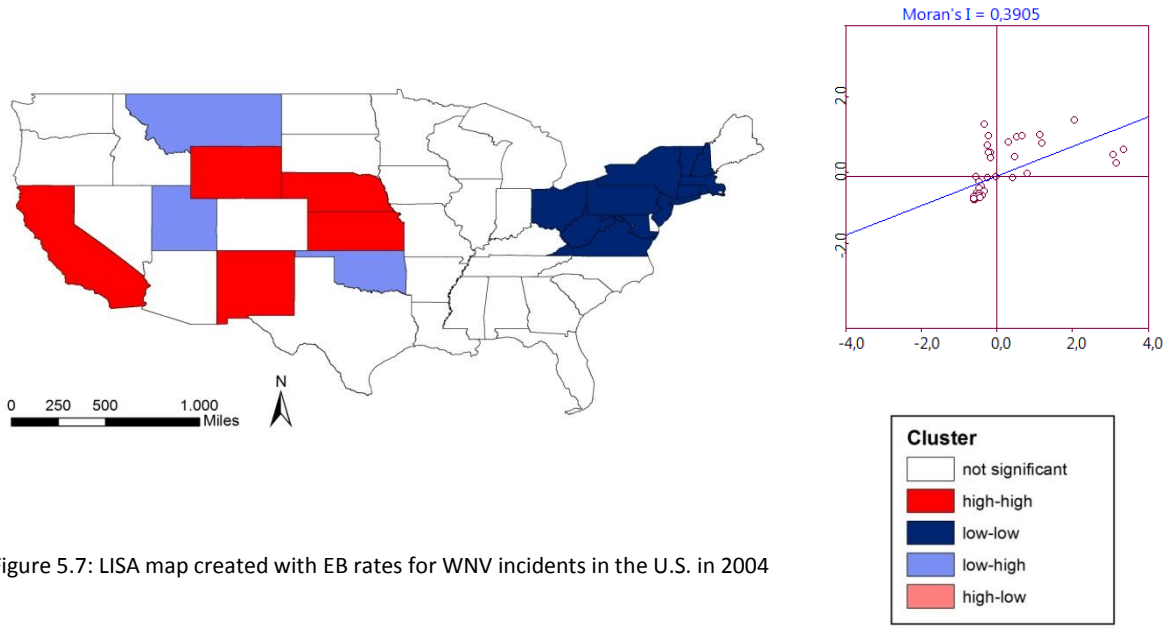


Figure 5.7: LISA map created with EB rates for WNV incidents in the U.S. in 2004

In 2005 the distribution is similar to the year 2003. There is a hot spot in the central and northern U.S. In the north-eastern U.S. there is a huge low-low cluster, indicating that WNV cases have been scarce in and around the dark blue shaded states.

2005

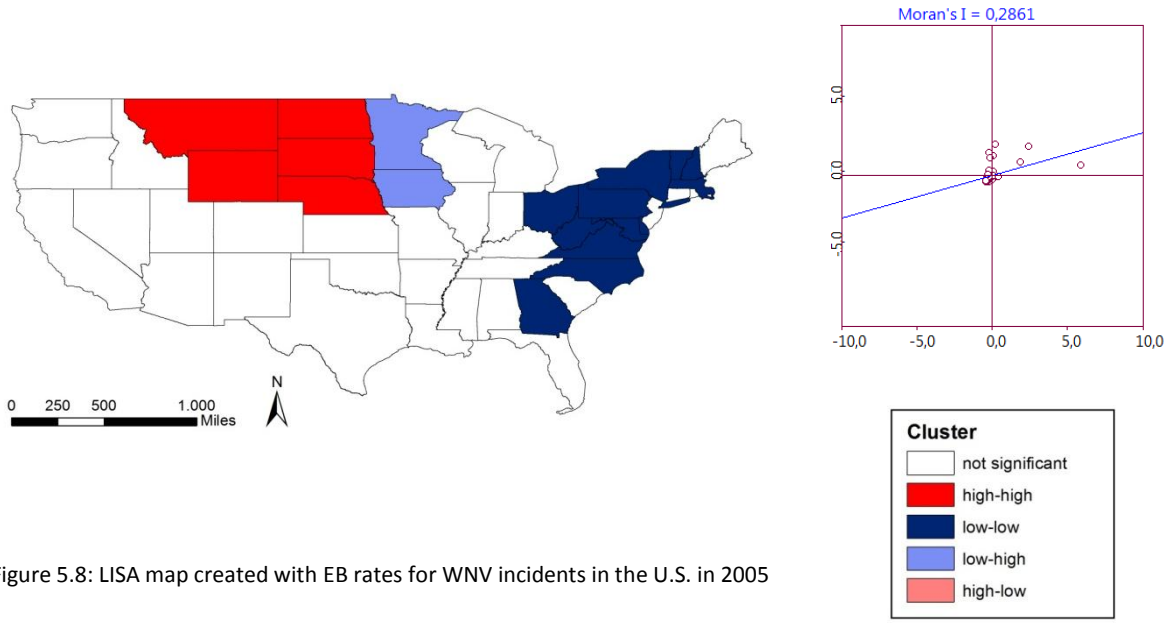


Figure 5.8: LISA map created with EB rates for WNV incidents in the U.S. in 2005

The WNV has become endemic to the U.S. by 2006. There is a high-high cluster in the west-central U.S., while in the north-eastern part hardly any cases occur; thus, the cold spot from the previous year persists. The Moran Scatter plot shows that the spatial autocorrelation approximates spatial randomness. One reason for this result is that by 2006 the WNV is present all across the U.S. and the reported cases are almost evenly distributed among on all states. Thus, not a lot of disease cases are concentrated in a particular region, rathermost cases are scattered randomly across the whole study area.

2006

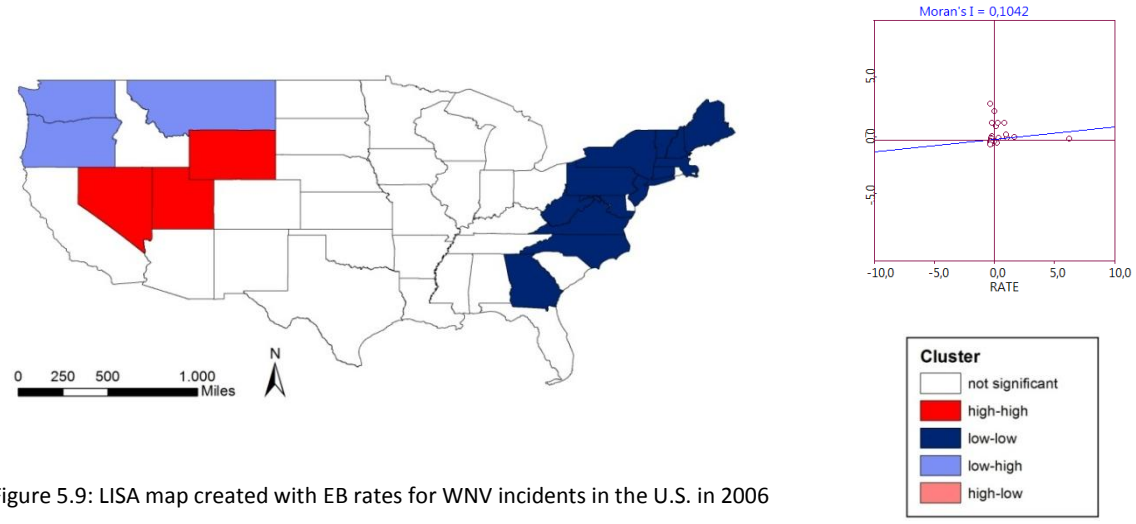


Figure 5.9: LISA map created with EB rates for WNV incidents in the U.S. in 2006

In 2007, the data distribution becomes more clustered again. There is one major hot spot in the central and northern U.S. Additionally, the cold spot on the east coast persists. The global spatial autocorrelation is positive (Moran's I = 0.4370).

2007

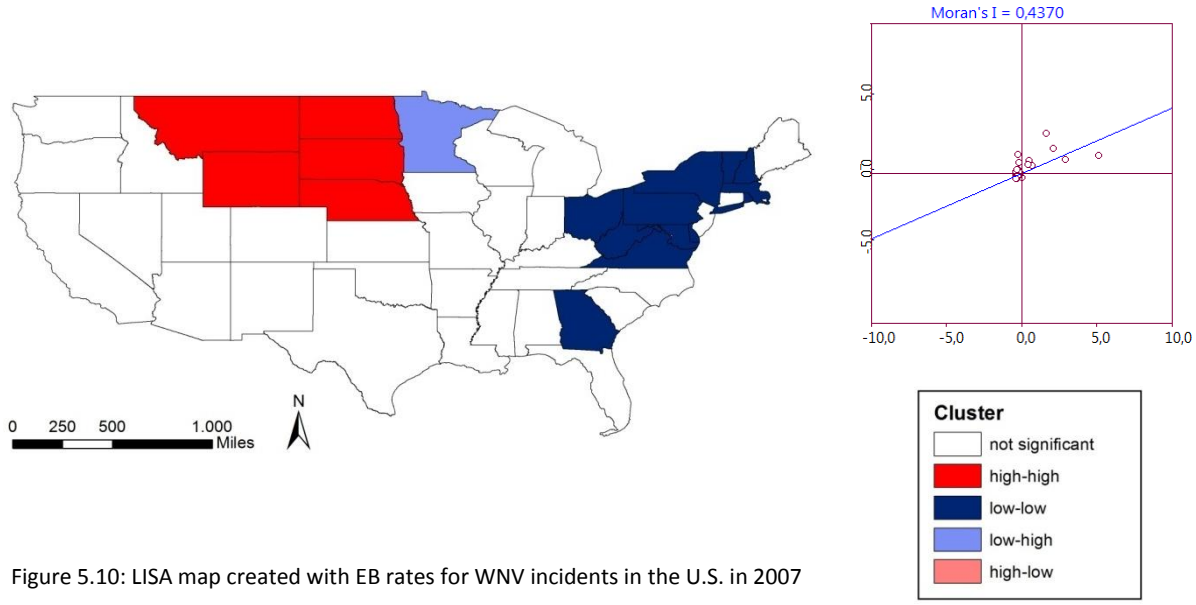


Figure 5.10: LISA map created with EB rates for WNV incidents in the U.S. in 2007

There are some minor differences in the year 2008 compared to the previous year. According to the global Moran's I there is only slight positive spatial correlation in the data, thus, the clustering effect is low. There continues to be still a hot spot in the central U.S. surrounded by states with low disease cases (light blue shaded polygons) and the omnipresent cold spot in the eastern U.S.

2008

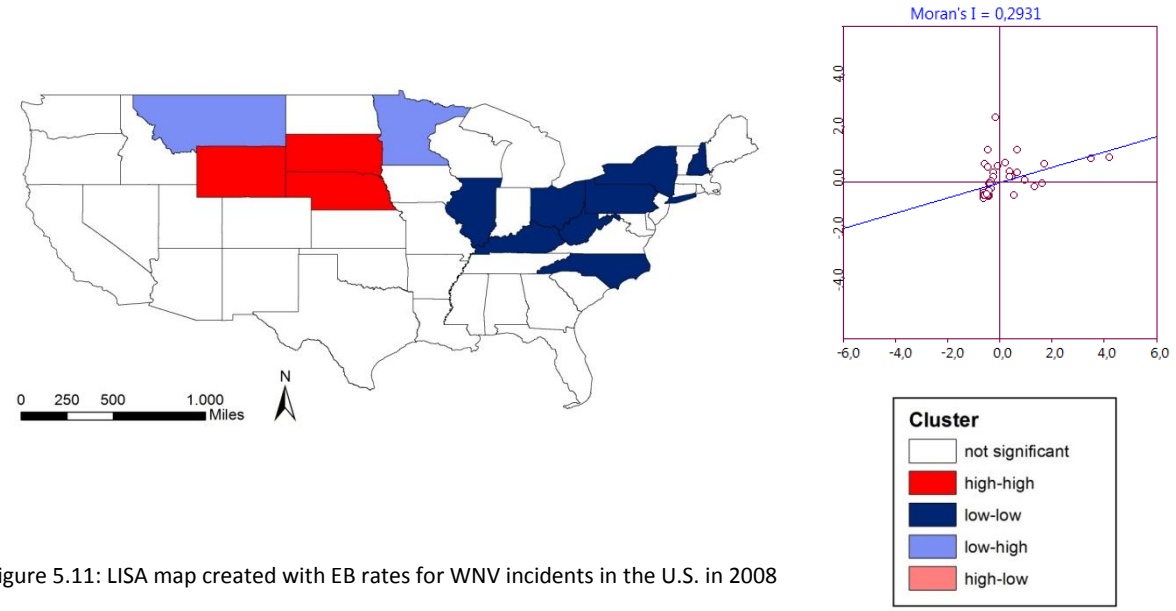


Figure 5.11: LISA map created with EB rates for WNV incidents in the U.S. in 2008

In 2009 the high-high cluster in the central U.S. has grown. There are two major clusters, including the cold spot in the east and the hot spot in the center of the contiguous U.S.

2009

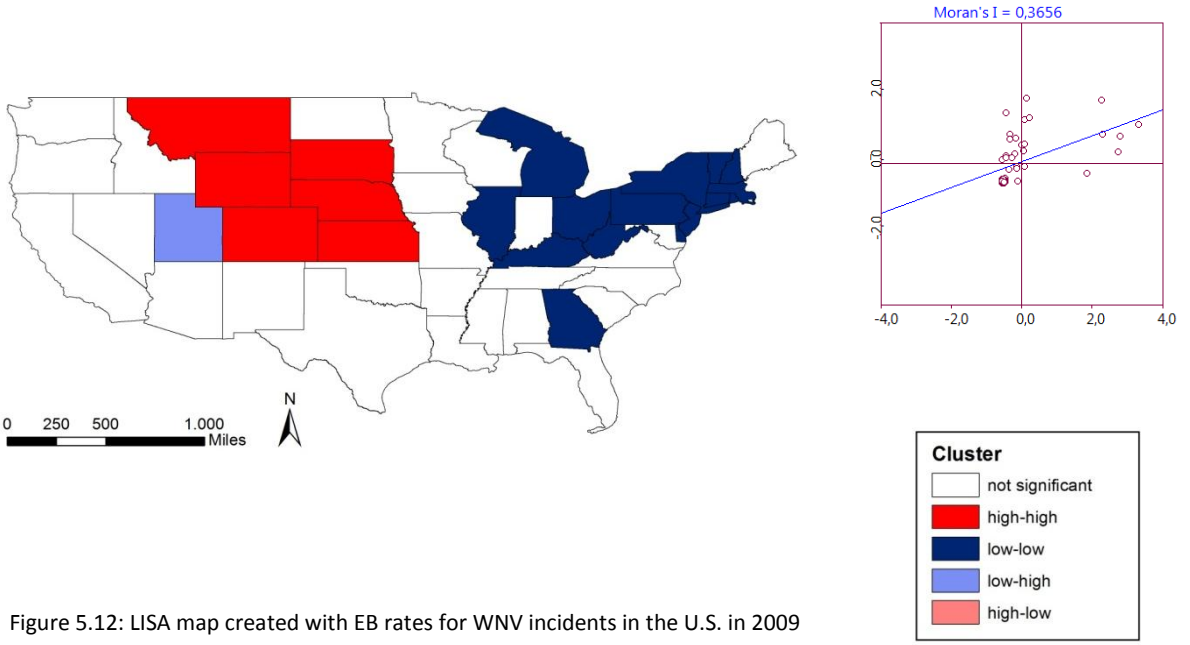


Figure 5.12: LISA map created with EB rates for WNV incidents in the U.S. in 2009

In 2010 the low-low cluster at the east coast has eventually disappeared. Instead, a cold spot has emerged in the south-east. One major high-high cluster is still present in the central U.S. In general, WNV incidents in the states have decreased steadily.

2010

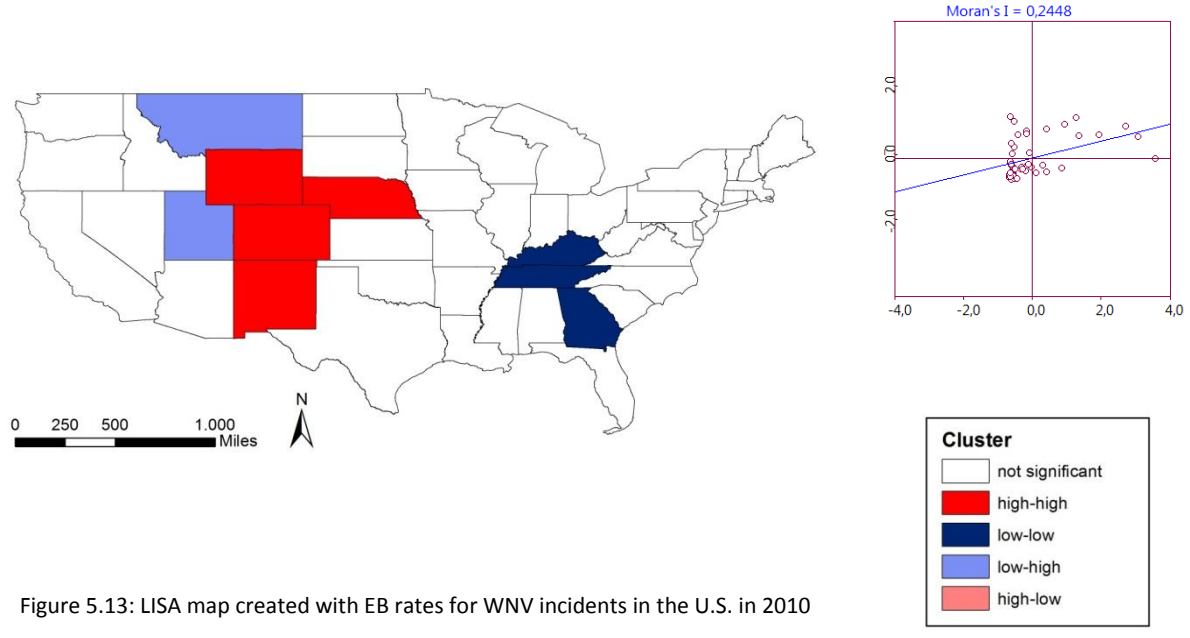


Figure 5.13: LISA map created with EB rates for WNV incidents in the U.S. in 2010

In 2011, most of the states are no longer significant as far as a clustering effect is concerned. The Moran's I is close to zero, which means that there is no spatial autocorrelation among the WNV cases across the states. This can be interpreted as a spatial random distribution.

2011

Figure 5.14: LISA map created with EB rates for WNV incidents in the U.S. in 2011

5.1.2 Retrospective and prospective test clustering in WNV disease data

The analysis for the U.S. in GeoSurveillance did not bring any results. This might be due to the input data, although several alternative U.S. shapefiles (e.g. census data, ArcMap data) have been tested. Thus, this section of the results stays empty.

5.1.3 Analysis of space-time clusters of WNV disease data using Kulldorff's Scan Statistic

The Scan Statistic can be applied to do retrospective and prospective analysis. Kulldorff (Kulldorff, 1997) defines retrospective analysis as a method to detect both "alive" clusters as well as "historic" clusters. The latter are clusters that ceased to exist before the end of the study period. For the retrospective analysis two different probability models, which comply best with the data selected for this study, have been chosen. The discrete Poisson probability model can be used if there are case data together with a background population at risk. The case data are included as part of the population count (Kulldorff, 1997). The second model is the Space-Time permutation model which is appropriate for use when case data are available. Both models are implemented for space-time analysis. SaTScan also provides options for purely spatial or purely temporal analyses. Prospective clustering methods are used for early detection of a disease outbreak. The analyses are conducted at continuous intervals, such as weekly, monthly, or yearly. For prospective clustering, the discrete Poisson probability model is applied, as well.

Using different probability models, retrospective analyses show slightly different results. The discrete Poisson model reveals two major spatio-temporal clusters which evolved throughout the investigation period, from 1999 to 2011 (see Figure 5.15). The most likely cluster is the red cluster labeled as “Cluster1”. It is the cluster that is least likely to be due to chance (Kulldorff, 2010). The cluster initially appeared in 2003 and ceased to exist in 2007. Altogether, there were 12,313 cases compared to 538 expected cases. This corresponds to a ratio of 23:1. The p-value is smaller than 0.001, thus the cluster is highly significant. The relative risk for the population inside the cluster compared to the population outside of the cluster is 36.98. This number is the estimated risk within the cluster divided by the estimated risk outside the cluster (Kulldorff, 2010). The summary of the analysis is provided in Figure 5.16. The analysis also reveals a secondary cluster (green shaded polygons). This cluster, however, only persisted over one period in 2002. The number of observed cases was 2534 compared to the expected cases of 471. This corresponds to a ratio of approximately 6:1. According to these calculations the relative risk inside the cluster is 5.76. The p-value, again, is smaller than 0.001.

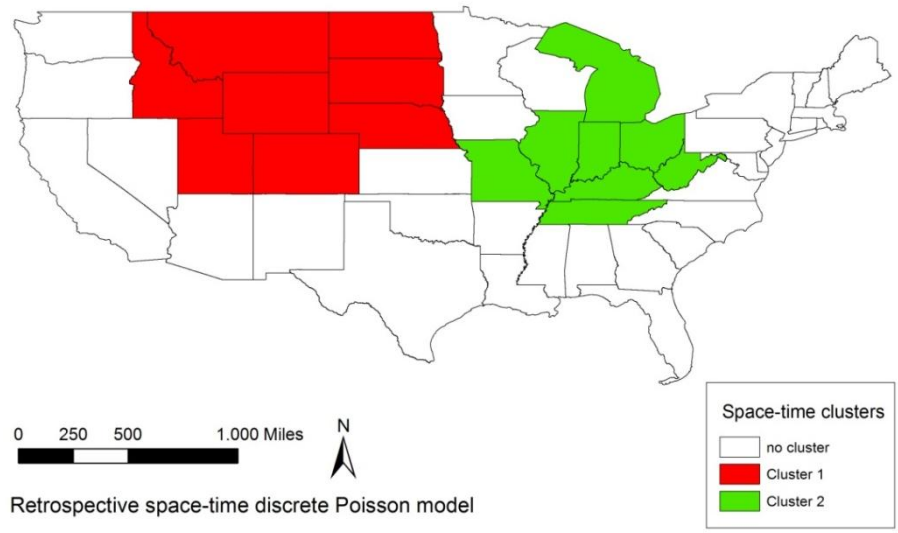


Figure 5.15: Retrospective test (discrete Poisson) for clustering in SaTScan

```
SUMMARY OF DATA

Study period...................: 1999/1/1 to 2011/12/31
Number of locations............: 49
Total population...............: 291571146
Total number of cases..........: 31389
Annual cases / 100000..........: 0.8

-------------------------------------------------------------------------------

MOST LIKELY CLUSTER

1.Location IDs included.: 56, 8, 30, 49, 16, 46, 31, 38
  Coordinates / radius..: (42.999627 N, 107.551451 W) / 742.04 km
  Time frame............: 2003/1/1 to 2007/12/31
  Population............: 13116096
  Number of cases.......: 12313
  Expected cases........: 538.49
  Annual cases / 100000.: 18.9
  Observed / expected...: 22.87
  Relative risk.........: 36.98
  Log likelihood ratio..: 29365.091052
  P-value...............: < 0.00000000000000010
```

```
SECONDARY CLUSTERS

2.Location IDs included.: 18, 17, 21, 39, 47, 54, 26, 29
  Coordinates / radius..: (39.919881 N, 86.281825 W) / 560.99 km
  Time frame............: 2002/1/1 to 2002/12/31
  Population............: 58027806
  Number of cases.......: 2534
  Expected cases........: 471.11
  Annual cases / 100000.: 4.5
  Observed / expected...: 5.38
  Relative risk.........: 5.76
  Log likelihood ratio..: 2270.858588
  P-value...............: < 0.000000000000000010
```

Figure 5.16: Summary of space-time analysis (discrete Poisson model) including results for the most likely and secondary clusters

Applying the space-time permutation model, one most likely spatio-temporal cluster (cluster 1) together with two secondary spatio-temporal clusters (cluster 2 & 3) is detected (see Figure 5.17). The most likely cluster is located in the eastern and central U.S., where the discrete Poisson model detected a secondary cluster. The time frame of the cluster, again, only extends over one year (2002). There are 2,586 observed cases compared to 627 expected cases. Thus, the ratio between observed and expected cases is 4.12. The p-value is smaller than 0.001, which indicates a high significance. Cluster 2 is a secondary spatio-temporal cluster where the observed/expected ratio is 1.92 with a p-value smaller than 0.001. This is a historic cluster only persisting in 2003. Cluster number 3 is also a historic cluster, which emerges and ceases in 2006. The observed/expected ration is 5.27. Since the p-value is smaller than 0.001, the cluster is highly significant (see Figure 5.18).

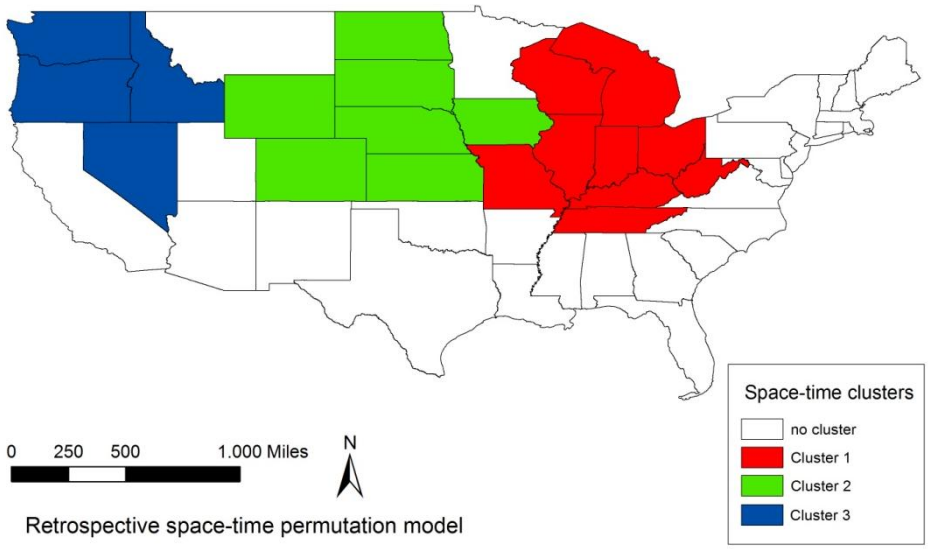


Figure 5.17: Retrospective test for clustering (space-time permutation) in SaTScan

```
SUMMARY OF DATA

Study period..................: 1999/1/1 to 2011/12/31
Number of locations...........: 49
Total number of cases.........: 31389

_____________________________________________________

MOST LIKELY CLUSTER

1.Location IDs included.: 18, 17, 21, 39, 47, 54, 26, 29, 55
  Coordinates / radius..: (39.919881 N, 86.281825 W) / 596.20 km
  Time frame............: 2002/1/1 to 2002/12/31
  Number of cases.......: 2586
  Expected cases........: 627.72
  Observed / expected...: 4.12
  Test statistic........: 1766.599547
  P-value...............: < 0.000000000000000010
```

```
SECONDARY CLUSTERS

2.Location IDs included.: 31, 46, 20, 19, 8, 56, 38
  Coordinates / radius..: (41.527151 N, 99.810856 W) / 659.83 km
  Time frame............: 2003/1/1 to 2003/12/31
  Number of cases.......: 7158
  Expected cases........: 3730.96
  Observed / expected...: 1.92
  Test statistic........: 1458.532362
  P-value...............: < 0.000000000000000010

3.Location IDs included.: 41, 53, 16, 32
  Coordinates / radius..: (43.940439 N, 120.605273 W) / 605.71 km
  Time frame............: 2006/1/1 to 2006/12/31
  Number of cases.......: 1189
  Expected cases........: 225.47
  Observed / expected...: 5.27
  Test statistic........: 1028.446626
  P-value...............: < 0.000000000000000010
```

Figure 5.18: Summary of space-time analysis (Space-time permutation model) including results for the most likely and secondary clusters

Interestingly, none of the retrospective models reveals any “alive” clusters. This might be a hint that the prevalence of WNV is ceasing. Thus, infections emerge more scattered across the 48 contiguous states, including D.C., to several illnesses before disappearing again. The information about distribution patterns of the WNV, clearly corresponds to the typical nature of the virus. First, it makes a big appearance in a particular region and subsequently spreads quickly to new areas. Depending on the aggressiveness of the viremia the virus

causes avian die-offs and several severe human disease cases. Then, it reaches a peak in terms of prevalence before ceasing again. However, the cessation of the virus can also be a result of awareness among people, precautionary measures, and comprehensive surveillance systems. For the latter, prospective cluster testing plays a big role in order to predict locations which are most prone in the future. In this analysis a prospective space-time discrete Poisson model has been applied to detect possible sites for clustering in the future. The most likely place, where the WNV could be a major issue in the future is Idaho. The cluster emerged in 2006 and continues to exist till the end of the study period in 2011. There have been 1,208 observed cases compared to 75 expected cases. This is a ratio of 16:1. The relative risk inside the cluster compared to locations outside the cluster is 16.68. A small p-value ($p<0.001$) confirms the high significance of the cluster. There are thirteen possibilities for secondary clusters. Most of them are located in the central or western U.S., in addition to the two southern states, Louisiana and Mississippi. Both states are potential WNV hot spots. Both have emerged in 2006 and are still "alive" at the end of the study period. In Louisiana the relative risk is 1.48, while in Mississippi the risk is 3.44. The map (see Figure 5.19) displays the most likely cluster (Idaho) as well as the thirteen secondary clusters. These are categorized on the basis of their relative risk ratio. The darker the shading, the higher is the relative risk. Among those states with a high risk ratio are Louisiana, Arizona, and Colorado.

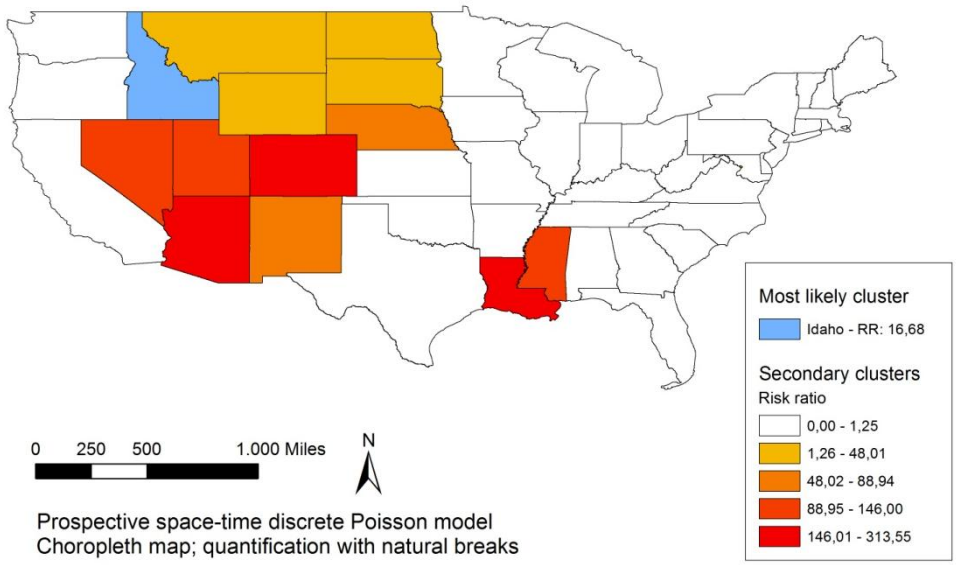


Figure 5.19: Prospective space-time clustering in SaTScan

5.1.4 Visualization of univariate space-time patterns (Vis-Stamp)

The next step in this analysis is to reveal the types of clusters that exist in the data. For this purpose, a SOM is applied in order to process the data and to derive clusters of spatial objects. The size for the SOM is chosen as 7x7, which is the default size in Vis-Stamp. That means that the SOM will be composed of 49 nodes which represent different types of clusters. Nodes can be left empty, thus no spatial object of the data is assigned to that kind of cluster. The SOM uses the Euclidean distance to assess similarity between spatial objects (Guo, 2009). When training the SOM, the program constructs a U-matrix with hexagons, representing the various cluster types (see Figure 5.20). The node hexagons contain circles, which are scaled on the basis of the number of data items belonging to the cluster. For

example, the large dark blue circle in the upper right corner of Figure 5.20 is a cluster type that contains a high number of states. In this specific case, the dark blue circle corresponds to states with less than one WNV incident among 10,000 people annually. This was the rate, which was used for the calculations. In the very opposite corner there is a dark red circle, representing a cluster type with approximately 1000 WNV cases per 10,000 inhabitants. The shading of the hexagons gives some indication about the dissimilarity between the clusters. This example, however, is more useful if the analysis is of multivariate nature. In this analysis dissimilarity is defined in terms of time and WNV intensiveness. Similar clusters are assigned similar colors from a 2D diverging color scheme (Guo, 2009).

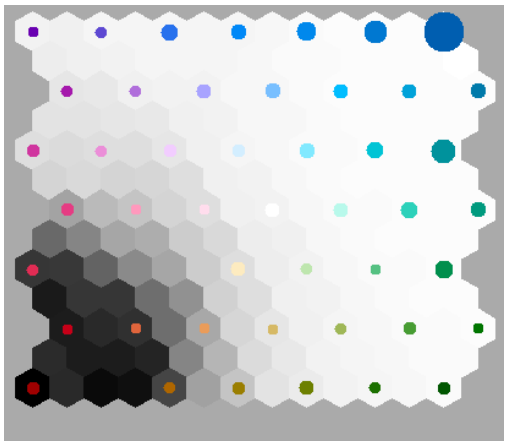


Figure 5.20: 7x7 SOM coloring clusters of spatial objects;

In the space-time matrix two dimensions are included, namely space and time. Thus, the analysis focuses on the space-time evolvement of the virus. The columns represent time (years) and the rows are the 49 contiguous U.S. states. While the temporal component follows a chronological order, the order of the spatial dimension seems to be random. That is not the case, because Vis-Stamp reveals similar clusters among the data and groups them. As it is depicted in Figure 5.21, in 2003 South Dakota, Nebraska, Nevada, Wyoming, and Colorado had significantly higher WNV incidents than the surrounding states. These five states are then considered to be a hot spot. The space-time matrix displays patterns among space-time data. In this analysis blue and green shaded objects correspond to a low WNV rate, whereas purple and red shaded objects are representative for a high WNV rate in a state. In the first year of the study period all states are shaded in blue because of the absence of the WNV in the U.S. New York State, however, is shaded in a lighter blue due to the initial WNV outbreak in the summer of 1999. Till 2001 there is no significant outbreak, but the virus spreads continually to adjacent states. Apart from 2002 the virus emerges in a high number of states and causes high infection rates. Thus, the hue of the matrix's cells goes from green to white and eventually reaches a luscious red. Earlier than ten years before its original appearance in the U.S. the WNV begins to cease gradually. There are hardly any high affected risk states any more. In 2011, most of the cells are already shaded in a blue hue again. According to the space-time matrix the most affected states in terms of intensity and longevity of the WNV have been South Dakota, Nebraska, and Nevada.

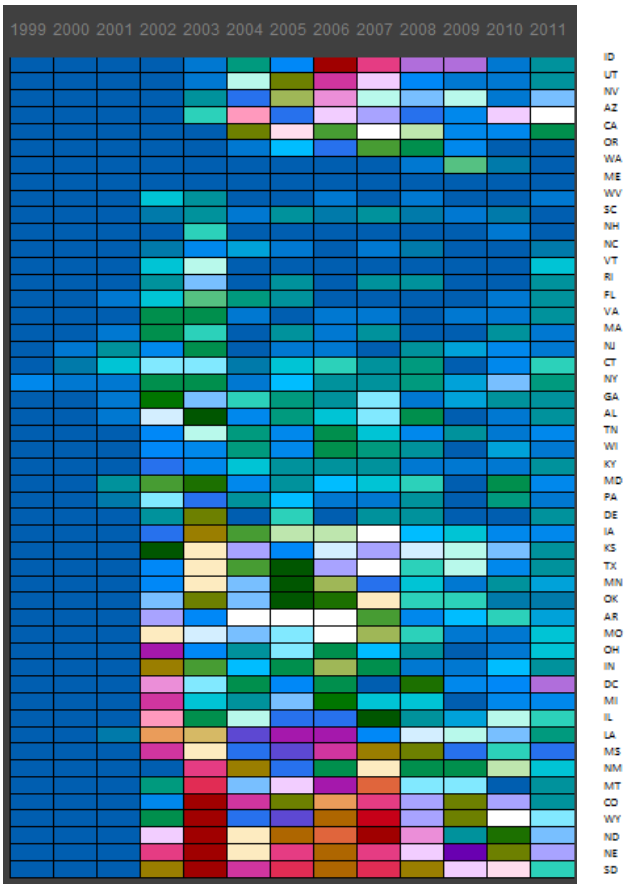


Figure 5.21: Space-Time Matrix

While the space-time matrix depicts each element, the map matrix highlights the entire study area year by year. Thus, changes in trends and patterns are easier recognizable. The legend for both the space-time matrix as well as the map matrix is a parallel coordinate plot (PCP) (see Figure 5.22). In this analysis the PCP only contains one variable. This is the rate of WNV disease cases per 10,000 people. A thick line corresponds to a high number of members in that cluster. A thin line corresponds to a low number of members. Similar clusters have similar colors. Blue hues are representative for a low number of cases, whereas red hues stand for a high number of cases in a state.

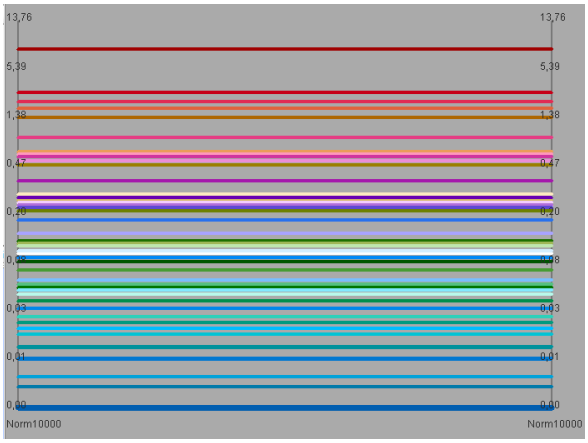


Figure 5.22: PCP – Parallel coordinate plot

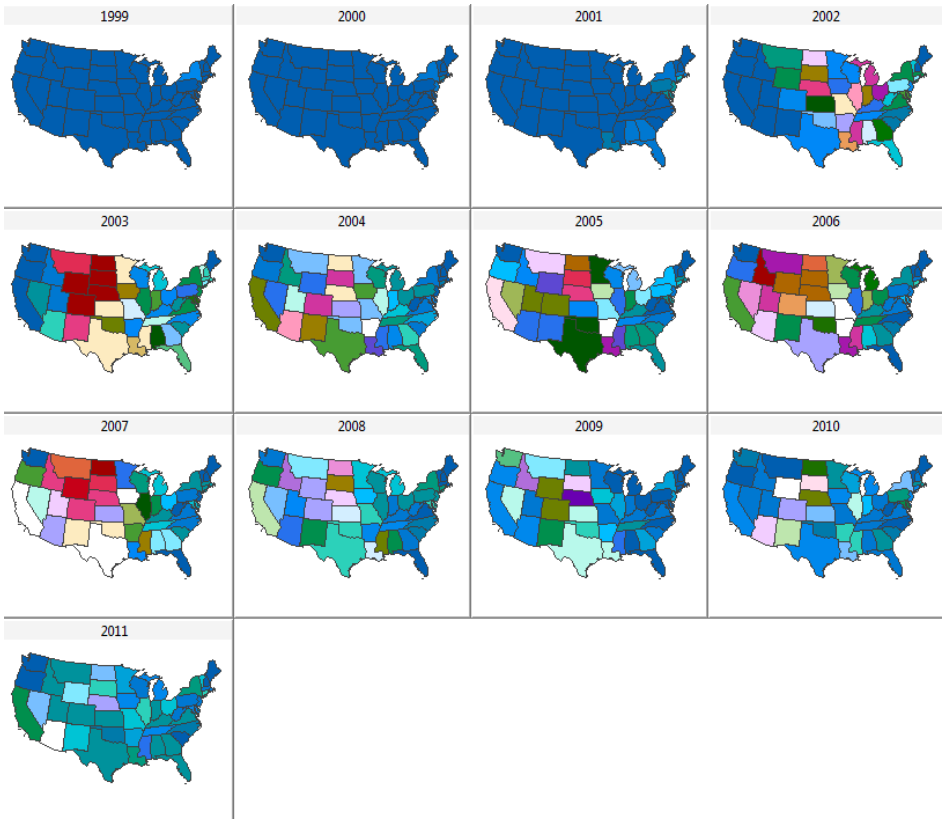


Figure 5.23: Map Matrix

In the map matrix (see Figure 5.23) the colors in the space-time matrix are projected onto the study area. In 1999 only the state of New York has a light blue hue. Until 2002, similar to the space-time matrix, there are only about three small distinct clusters in the study area. Each of those clusters holds few or no WNV cases. In 2002, the situation changes since the WNV spreads to the center of the U.S. The disease cases in several states rise significantly in comparison to the previous year. In 2003, a hot spot has evolved in the central U.S. In the states which belong to that cluster, the WNV disease rate is between 6 and 14 cases per 10,000 inhabitants. This is the highest rate occurring in the study period. In the following years, the virus continues its spread across the entire contiguous U.S., hitting California more severely in 2005. While several similar hot spots evolve in the central and western U.S., the north-eastern part of the U.S. goes back into a state with a low number of WNV incidents. Thus, there are hardly any WNV incidents in 2004 till the end of the study period in the north-east. The WNV has been most “successful” in terms of spread over the country in the years 2002 till 2007. This is when the virus “conquered” large areas and led to high disease rates in several states. In 2011, the last year of the study period only the states of Nebraska, Arizona and California reveal slightly higher WNV disease cases than the rest of the contiguous states of the U.S.

5.2 Spatial and Temporal analysis of the WNV in Louisiana

After the analysis of WNV distribution patterns in the U.S., the focus is now on the emergence and evolvement of the WNV in the state of Louisiana. The WNV initially emerged in Louisiana in 2001 with one registered human case (CDC, 2011a), and spreading continually across the entire state in the following years. In general, the transmission period is from June to November (CDC, 2011a). During this period most WNV disease cases are diagnosed and registered.

5.2.1 Exploratory spatial data analysis (ESDA) in Open GeoDa

Similar to the contiguous U.S., the first step in the analysis is to find out if there is spatial autocorrelation among the WNV cases and if what type of spatial autocorrelation exists, if any. For this purpose LISA maps are created in Open GeoDa. These maps highlight local spatial autocorrelation. For a global overview the univariate Moran's I is calculated. To measure spatial autocorrelation a weights file has to be created first. All spatial autocorrelation calculations are based on the Rook contiguity. The connectivity chart gives information about the neighborhood situation among Louisiana Parishes (see Figure 5.23). Thus, in the state of Louisiana most parishes have between four and six neighbors. For the creation of LISA maps Empirical Bayes rates are used. The total number of WNV incidents in a parish is normalized by the resident population estimate of the year 2002. The population estimate for Louisiana is provided by the U.S. census bureau.

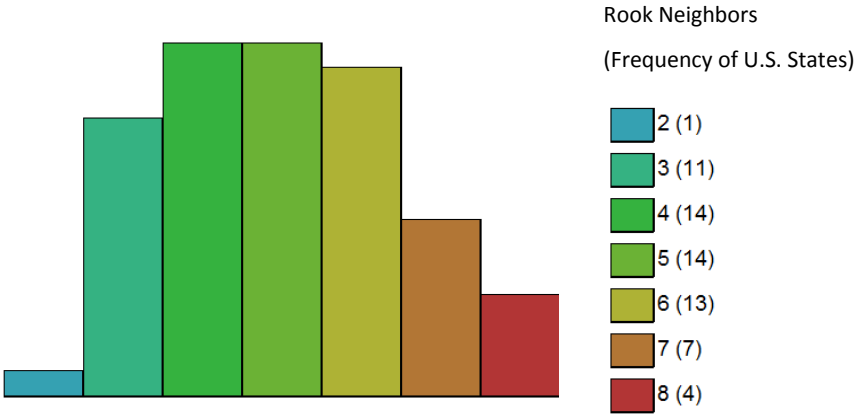
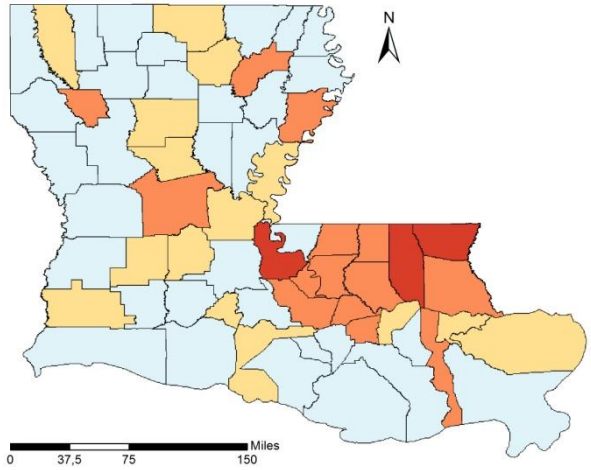
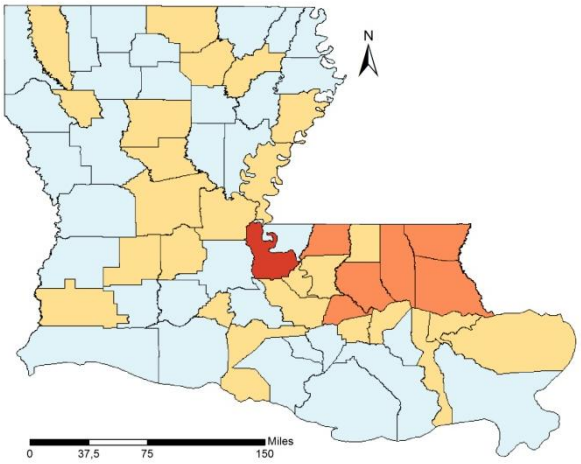
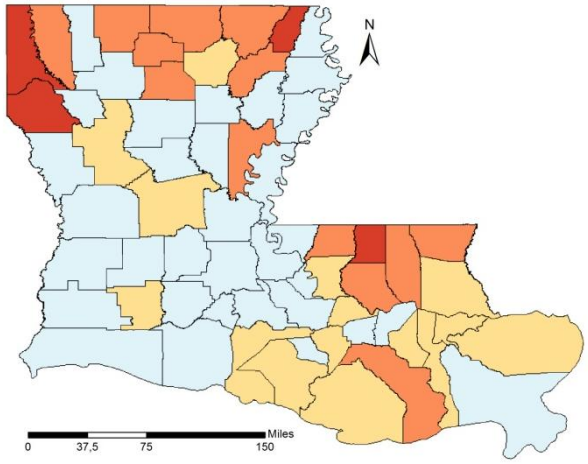
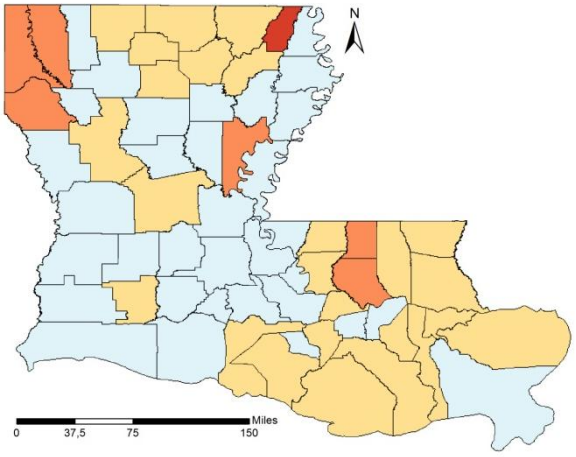
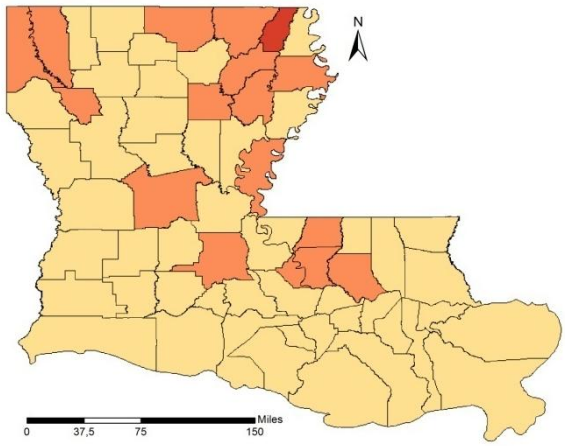
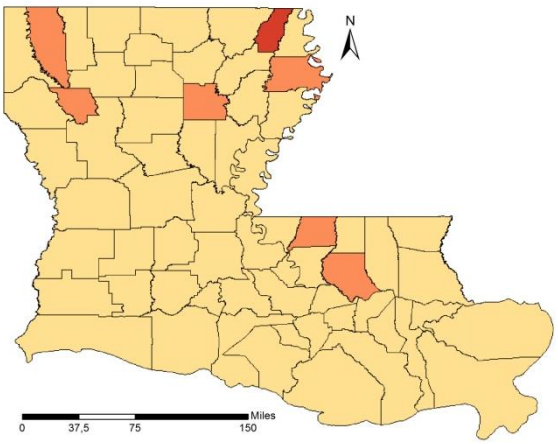
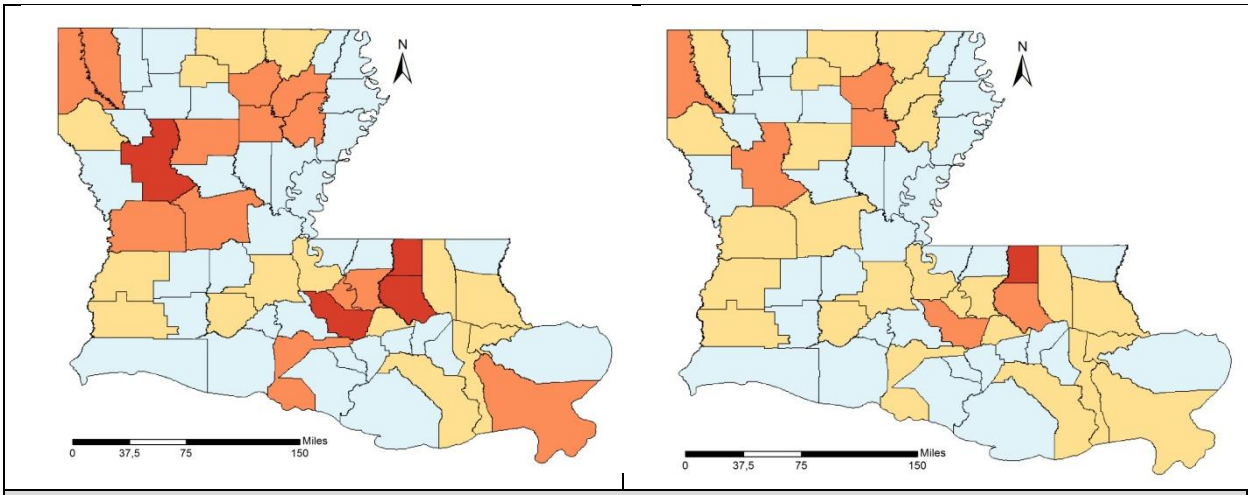


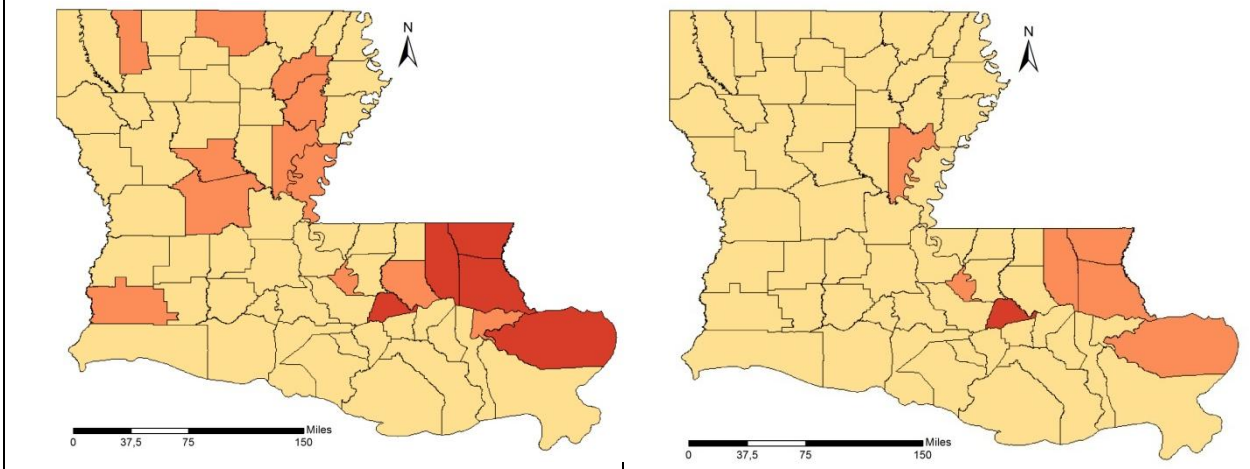
Figure 5.24: Connectivity chart for Louisiana parishes

First, lower and upper outliers are determined in order to get a better understanding of the WNV rates. The "raw rate" is a simple ratio of the event count to the base population at risk, while the Empirical Bayes procedure particularly affects the value for locations with small populations at risk (Anselin, 2003b). It will also typically remove the problem associated with many ties (especially zero values) (Anselin, 2003b). For this purpose, both Box Plot Maps as well as Percentile Maps are created with raw rates. The total number of disease incidents is normalized with the population estimates for the year 2002, provided by the U.S. census bureau.

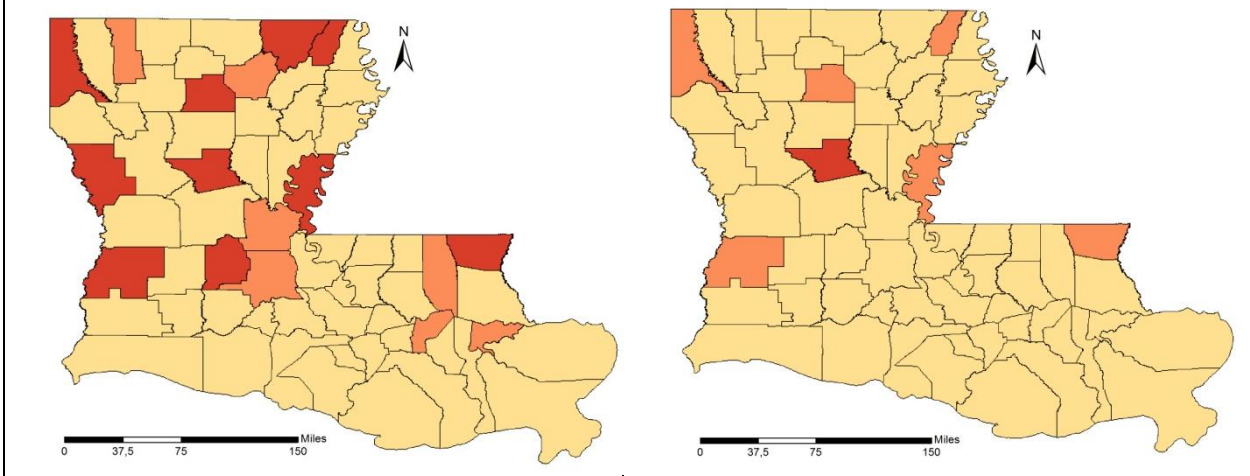
Box Plot Map Smoothed with raw rates Hinge = 1.5	Percentile Map Smoothed with raw rates
2002	
	
2003	
	
2004	
	
2005	



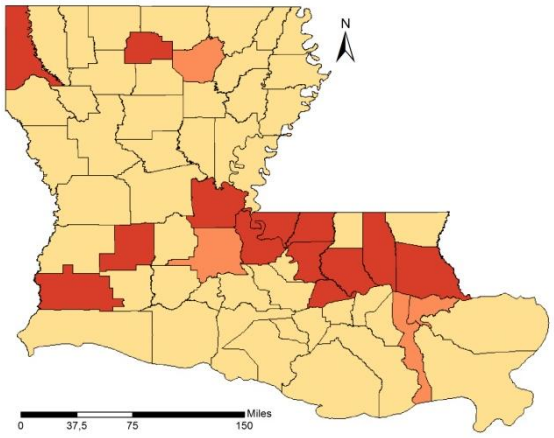
2006

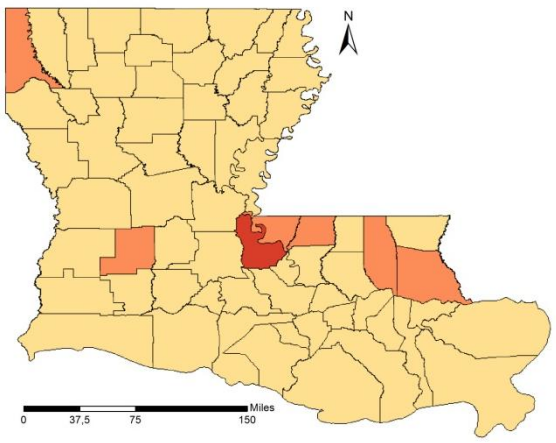


2007

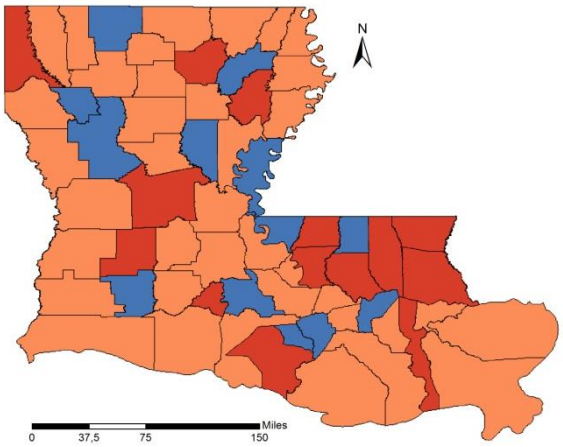


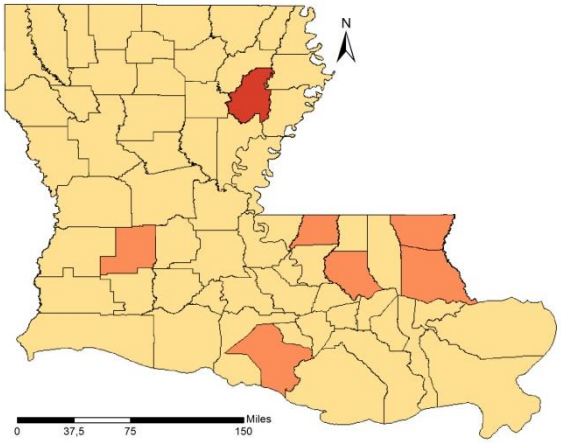
2008



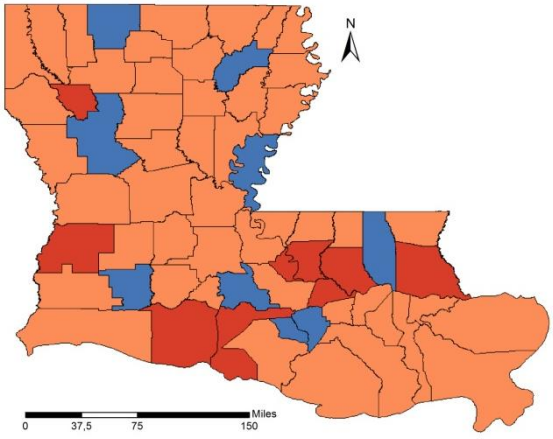


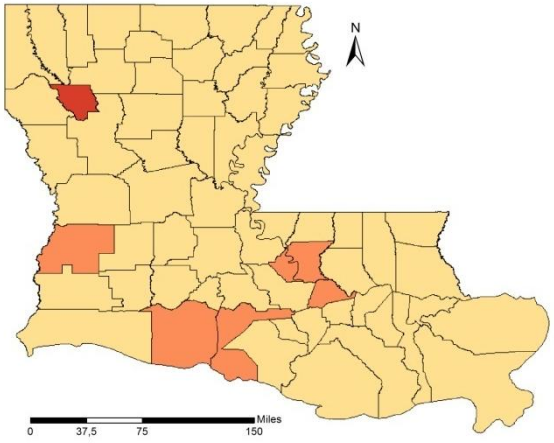
2009





2010





2011

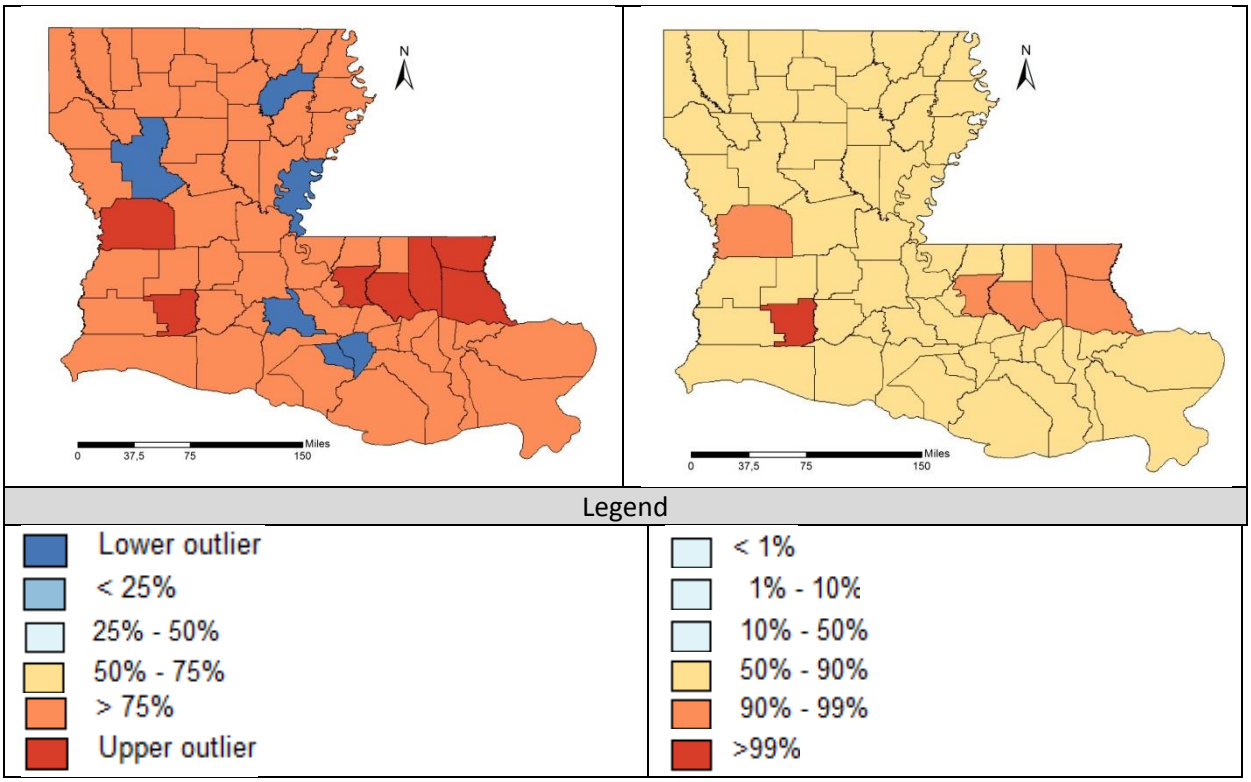


Table 5.2: Outlier maps of WNV disease cases in Louisiana in the study period from 2002 - 2011

The outlier analysis provides information about the location of lower and upper outliers in the WNV data. In 2002, there was a sudden WNV outbreak. Most of the upper outliers are located in the eastern part of the state, close to Jefferson Parish where the WNV was initially isolated in 2001. In the following year (2003), a change occurred. Several upper outliers appear in the very northern parishes of the state. The parishes in the east are still categorized in the upper quartile or in the upper 50% respectively. In 2004 the WNV spread across the entire state, leading to many similar count cases in the parishes. Apart from 2005, the upper outliers switch between the eastern part and the northern part of Louisiana. In 2008, most of the upper outliers are concentrated around the Baton Rouge metropolitan area. These outliers persist till the end of the study period in 2011, confirming the theory that the WNV is an urban disease. This means that the WNV finds perfect conditions to flourish in an urban environment. In 2009, for the first time some lower outliers appear. The lower outliers are scattered all over Louisiana. In 2011, again, most of the upper outliers are located in east Louisiana.

In Louisiana the WNV was initially only found in Jefferson Parish in August 2001 (see Figure 5.25). The first indication of the WNV presence was found in a crow. A little later, the first human infection was detected in a homeless man. In 2002, there was a large WNV outbreak, leading to 204 neuroinvasive disease cases (DHH, 2002). In addition, 104 cases with WNV fever have been reported. The outbreak started in mid-June, reaching a peak in July and August before the virus started to cease again in mid-August.

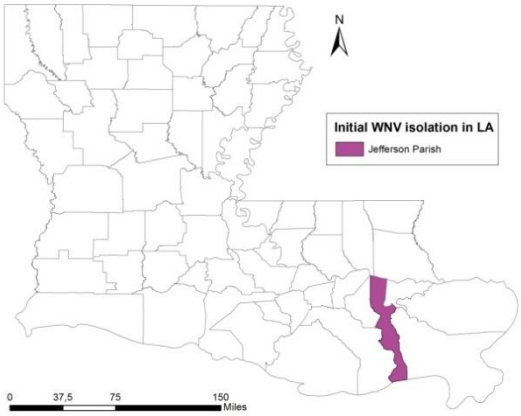


Figure 5.25: Introduction of the WNV in Jefferson Parish in 2001

The global Moran's I reveals a slight positive spatial autocorrelation, which is an indicator for clustering. Indeed, there is a hot spot north of Jefferson Parish. Two low-low clusters are located at the coast of Louisiana (see Figure 5.25).

2002

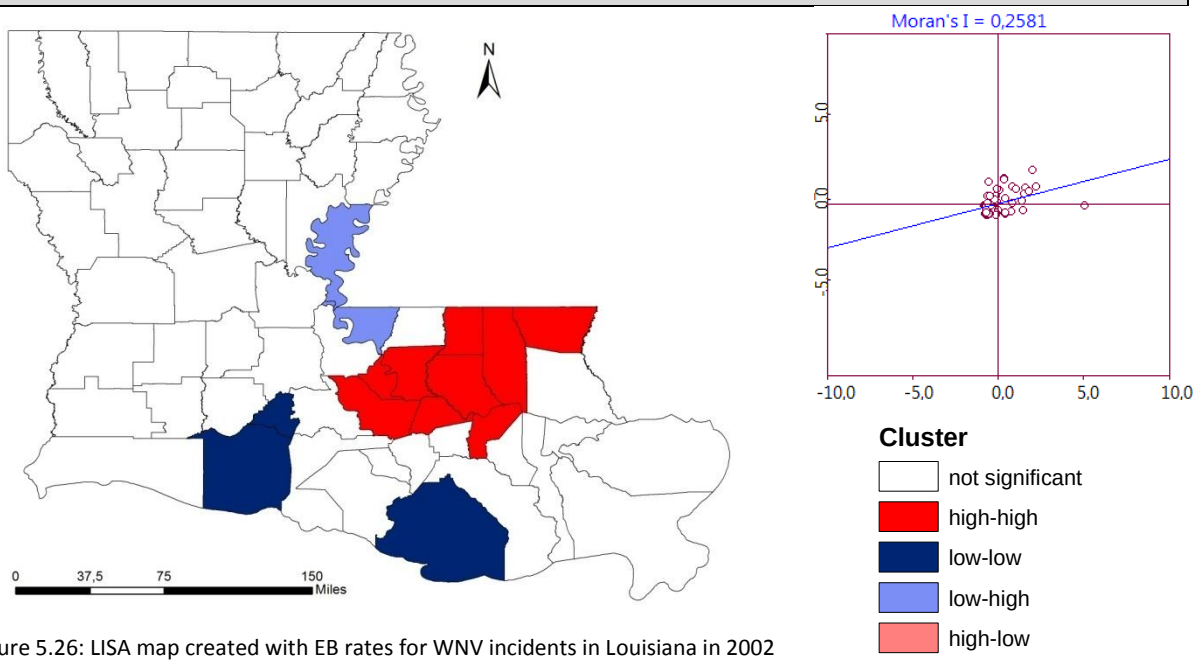


Figure 5.26: LISA map created with EB rates for WNV incidents in Louisiana in 2002

In 2003, the hot spot in the southeastern part of Louisiana was dissolved. Curiously, there are two high-high clusters in the very northern part of the state. In central Louisiana a cold spot has evolved including the parishes Acadia, Lafayette, Pointe Coupee and St. Landry. In 2003, the number of human disease cases increased sharply but at the same time the duration of the arboviral season lengthened (DHH, 2003).

2003

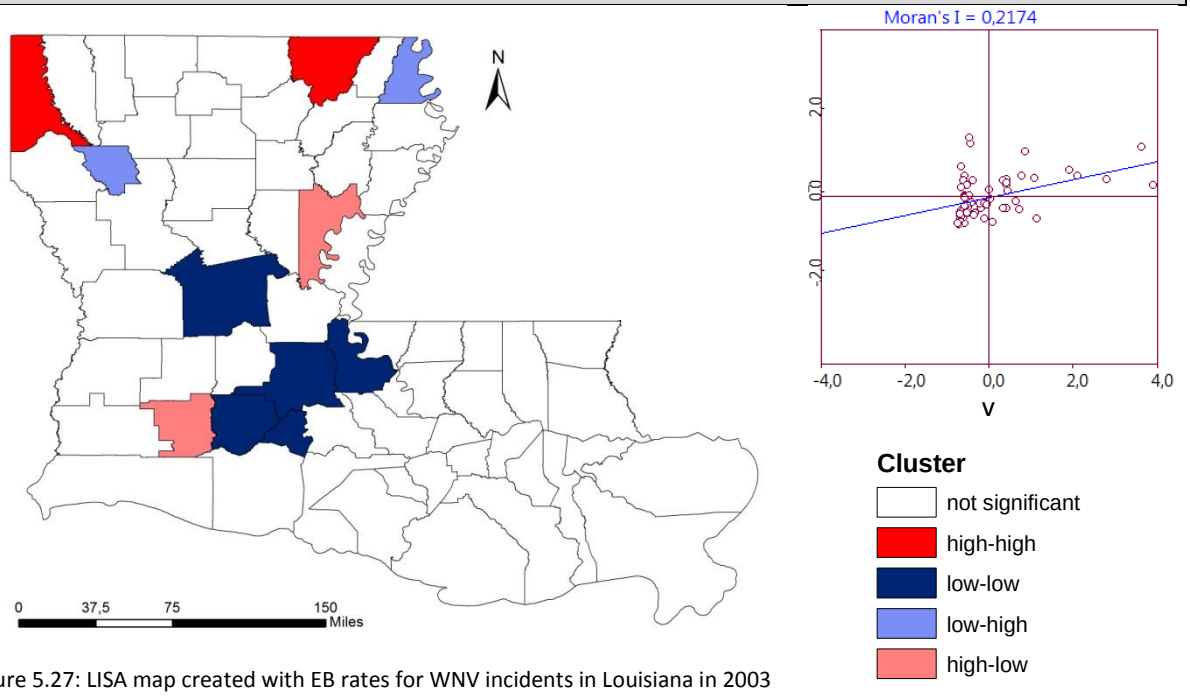


Figure 5.27: LISA map created with EB rates for WNV incidents in Louisiana in 2003

In 2004, a large low-low cluster has emerged in southern Louisiana, nearly covering the entire coastline of the state. The number of cases in the state still remains low. The center of most human activity in terms of virulence is the Baton Rouge metropolitan area extending over East Baton Rouge Parish and Livingston Parish (DHH, 2004). In addition there is a hot spot in the north, which has already emerged during the previous year. The high-low cluster in Rapides Parish, in central Louisiana, suggests that its WNV disease rate is higher than the rates in the adjacent parishes.

2004

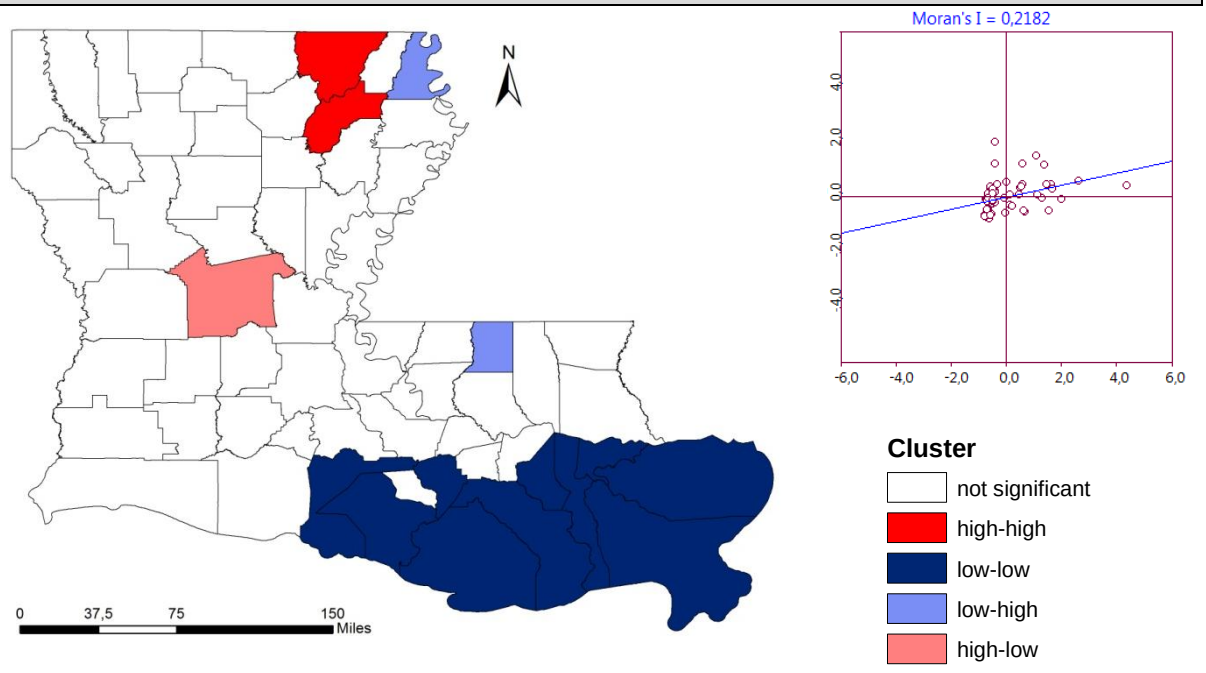


Figure 5.28: LISA map created with EB rates for WNV incidents in Louisiana in 2004

The Moran Scatter plot reveals that the Moran's I approaches spatial randomness. Still, many parishes are located in the third quadrant (lower-left quadrant) of the scatter plot, indicating the occurrence of low-low clusters. There is still one cold spot at the Gulf Coast. In the north of that cold spot is a high-high cluster, including East Baton Rouge Parish. The number of total cases (183) is higher than in the last two arboviral seasons, but still lower than during the outbreak in 2002.

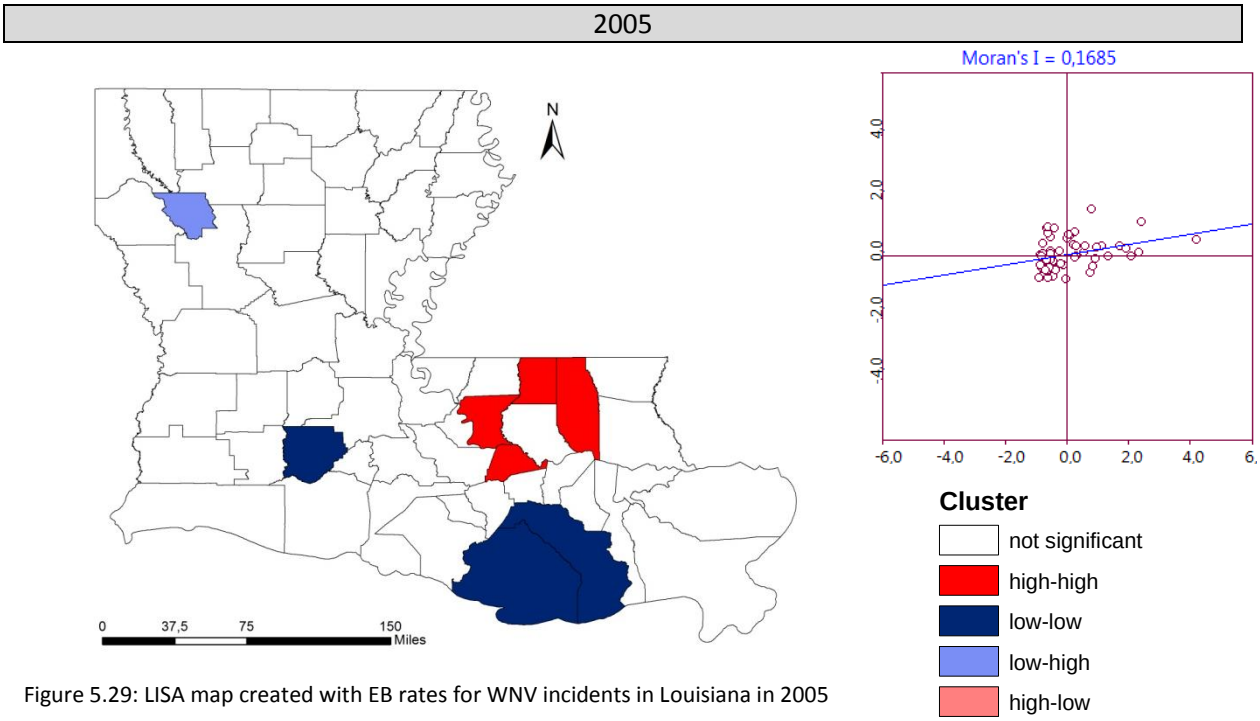


Figure 5.29: LISA map created with EB rates for WNV incidents in Louisiana in 2005

In 2006, the spatial distribution does no longer suggest any spatial autocorrelation (but this is not true, since the Moran's I is 0.2472). Thus, there is no dependency between the WNV rates of neighboring parishes (again, this is not really true). In addition, the parishes seem to be uniformly arranged around the origin in the Moran Scatter plot. However, the first quadrant is an exception, because the data are more scattered in that quadrant. There is a hot spot extending eastward from East Baton Rouge Parish to St. Tammany and Washington Parish. Within this hot spot disease rates are significantly higher than in the rest of the state. This is the reason for the data scattering in the first quadrant, where several parishes lie far away from the origin. In 2006 the number of disease cases is a little higher than in the previous years.

2006

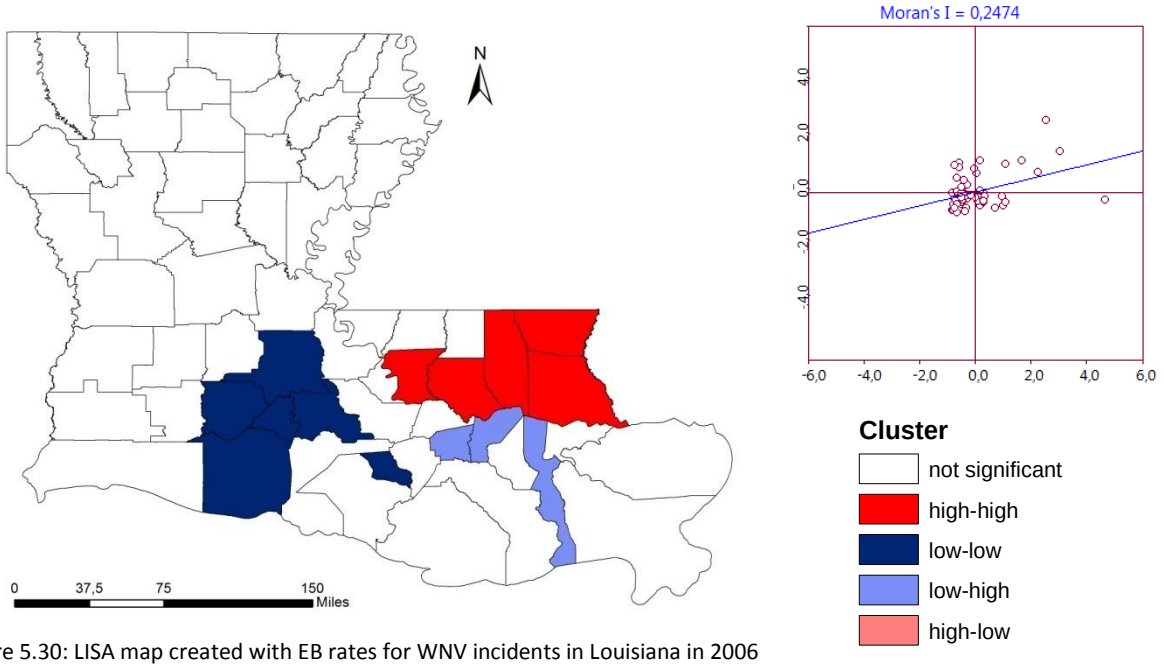


Figure 5.30: LISA map created with EB rates for WNV incidents in Louisiana in 2006

In 2007, the WNV infection rate has rapidly decreased. The global Moran's I still suggests a close to random distribution with a very weak hint of negative spatial autocorrelation. There are no more hot spots but there is one large cold spot at the Gulf Coast.

2007

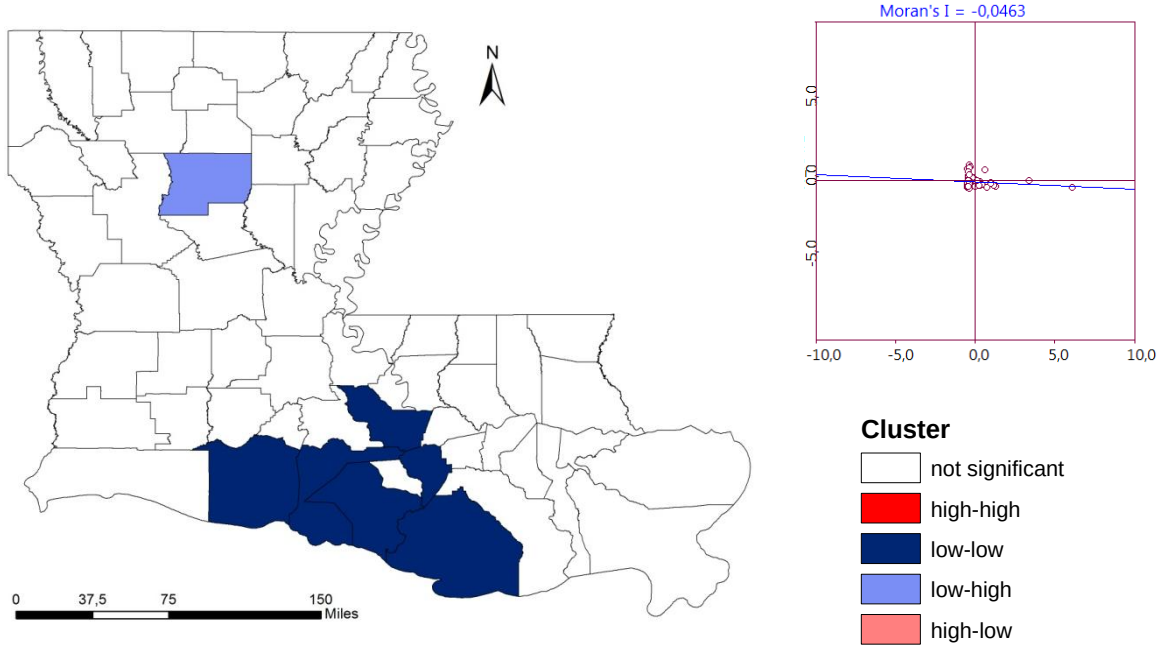


Figure 5.31: LISA map created with EB rates for WNV incidents in Louisiana in 2007

In 2008 WNV-related disease cases are already rare. Most of the incidents concentrate in central Louisiana, which are responsible for the creation of two hot spots in West Feliciana Parish and Avoyelles Parish. Along the Gulf Coast there are hardly any reported cases, resulting in a large cold spot.

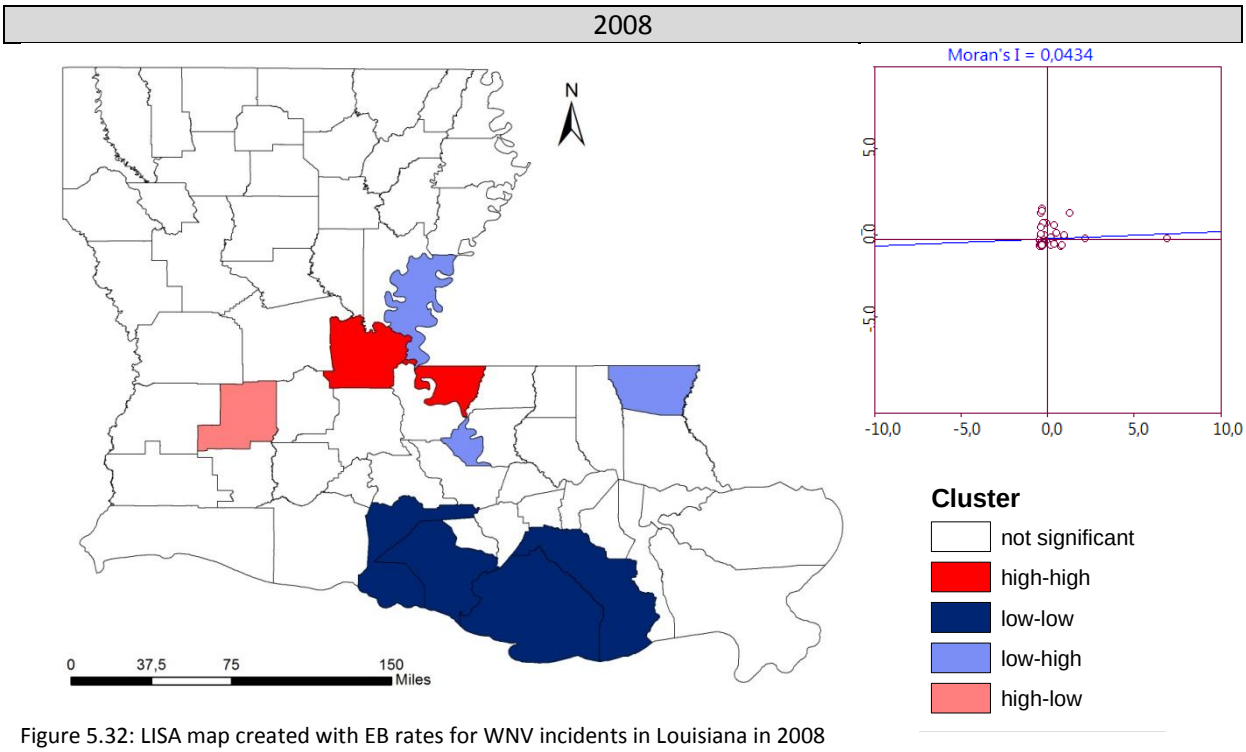


Figure 5.32: LISA map created with EB rates for WNV incidents in Louisiana in 2008

The spatial distribution of WNV rates from 2009 is slightly different compared to the previous years. Most parishes are concentrated in both the second (upper -left) as well as the third (lower-left) quadrant. While the second quadrant is an indicator for the occurrence of spatial outliers, the third quadrant stands for the existence of cold spots. Indeed, there are several low-low clusters in the southern part of Louisiana. In addition there are two hot spots in eastern Louisiana. The total number of disease cases, however, keeps getting lower.

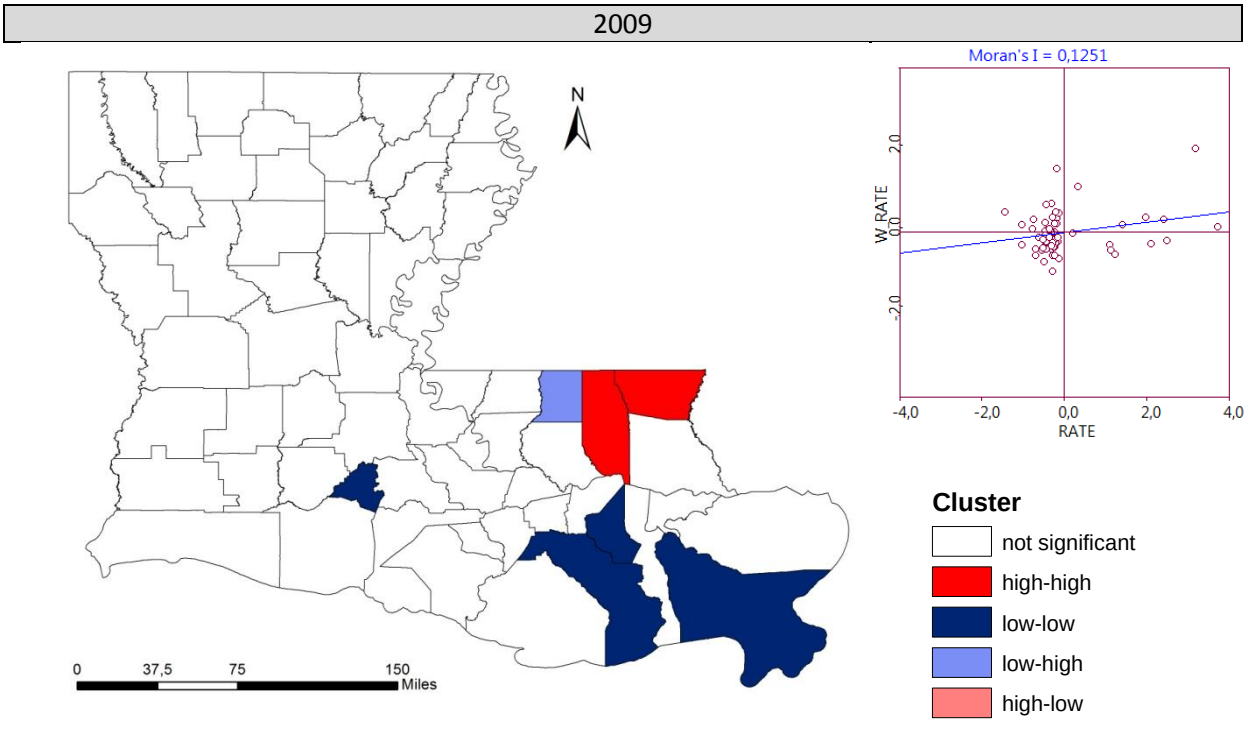


Figure 5.33: LISA map created with EB rates for WNV incidents in Louisiana in 2009

In 2010, there is a hot spot including East Baton Rouge Parish and Livingston Parish. Thus, most WNV activity in the arboviral season in 2010 is recorded in the Baton Rouge

metropolitan area. Several cold spots are scattered around the hot spot. No clusters or spatial outliers are found in the northern Louisiana region.

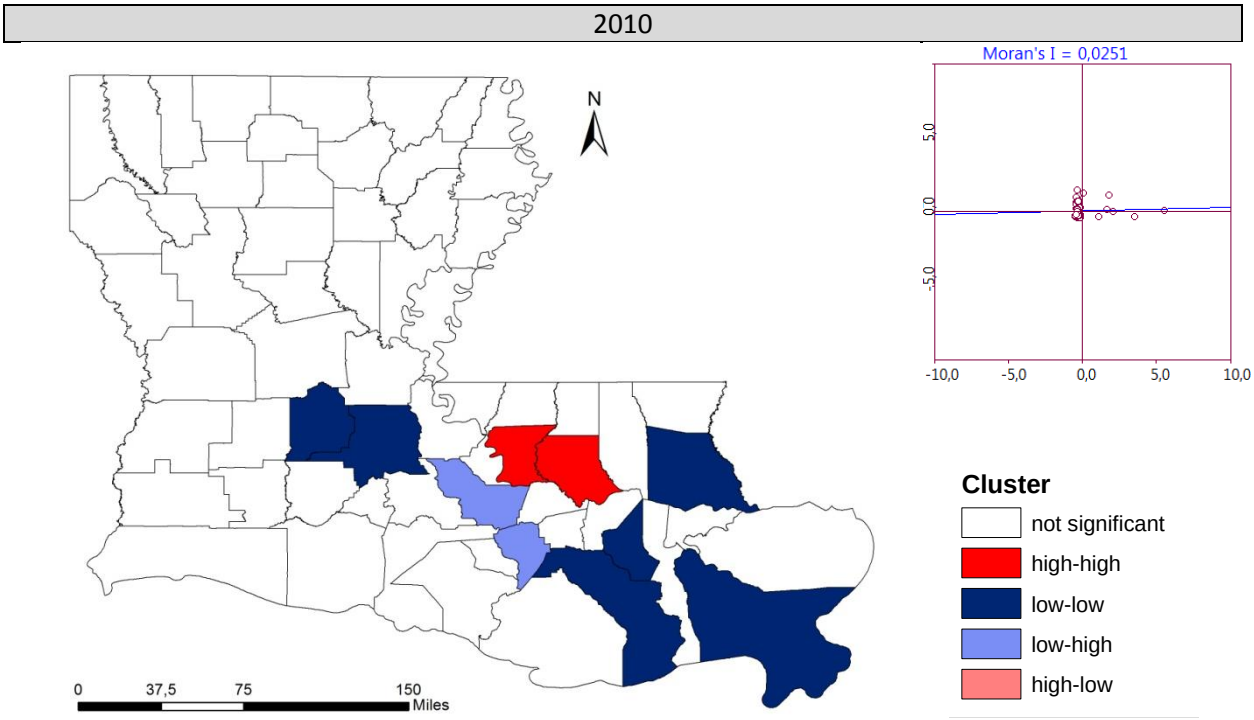


Figure 5.34: LISA map created with EB rates for WNV incidents in Louisiana in 2010

In 2011 WNV surveillance has faced an unprecedented low disease rate since its onset ten years earlier. During the whole season there have only been 12 reported human disease cases. Most of the cases are reported from eastern and southern Louisiana.

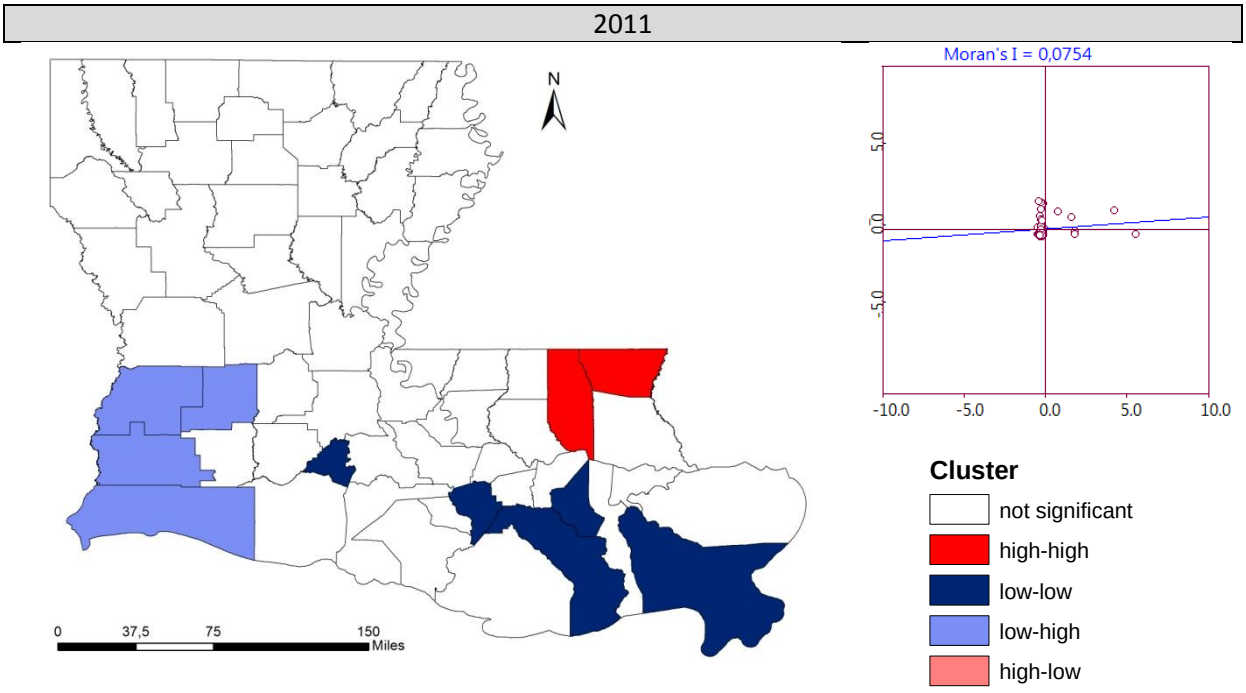


Figure 5.35: LISA map created with EB rates for WNV incidents in Louisiana in 2011

5.2.2 Retrospective and prospective tests for clustering in WNV disease data

Unlike Open GeoDa, GeoSurveillance has the power to integrate a temporal dimension into analysis. Thus, one single test is conducted in order to detect clusters which have emerged in Louisiana and are still persistent at the end of the study period. This methodology is called prospective testing and is implemented when there is a need to discover emerging or prospective clusters. Retrospective tests are conducted for every single observation year from 2002 till 2011.

5.2.2.1 Retrospective tests

Retrospective tests are calculated with the adjusted Poisson model, considering a fixed bandwidth and a significance level α = 0.05. In the legend (see Figure 5.36) the local score statistic U is compared to predefined critical threshold values. In the case of the score statistic, category endpoints are defined as zero, one-third and two-thirds of the maximum and minimum values (Rogerson et al., 2007).

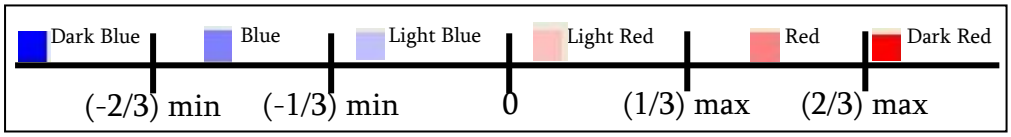
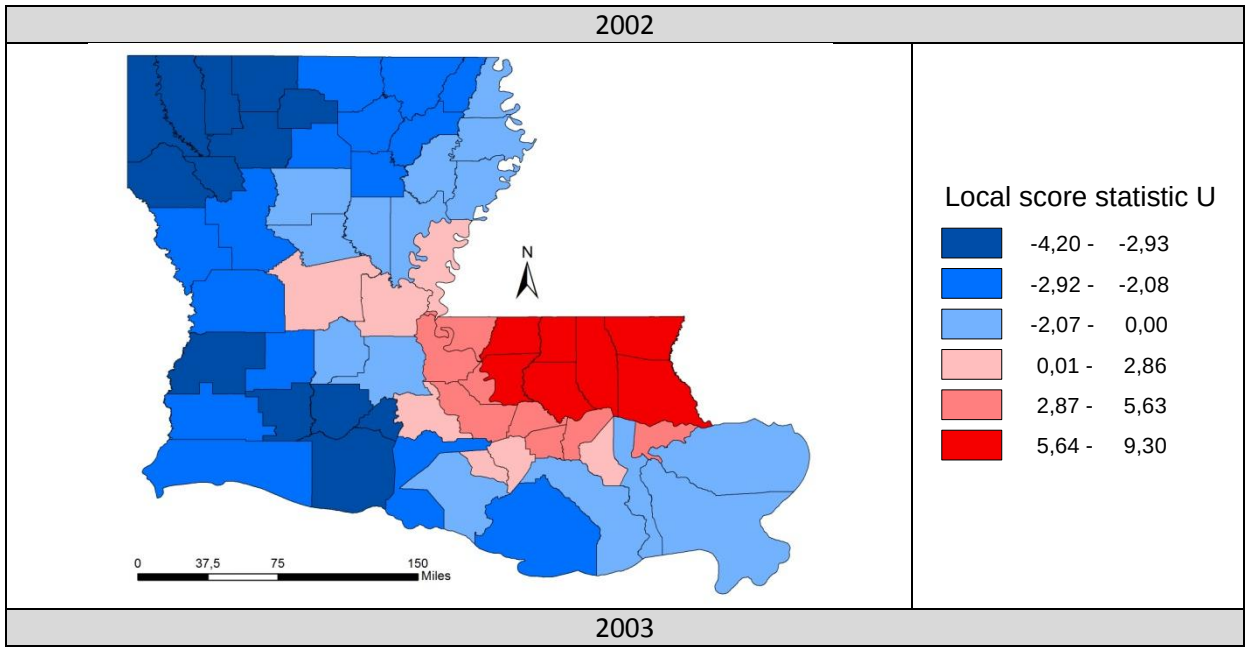
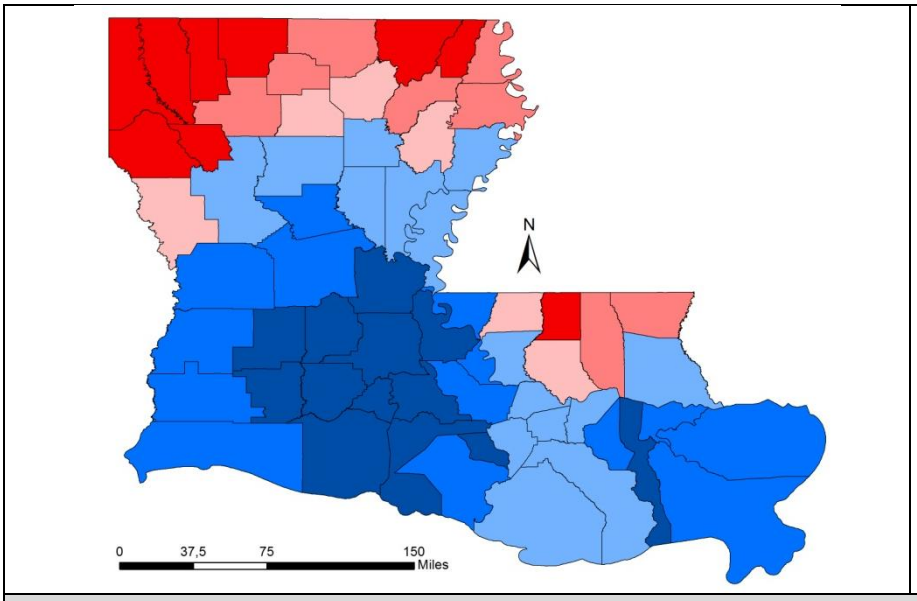
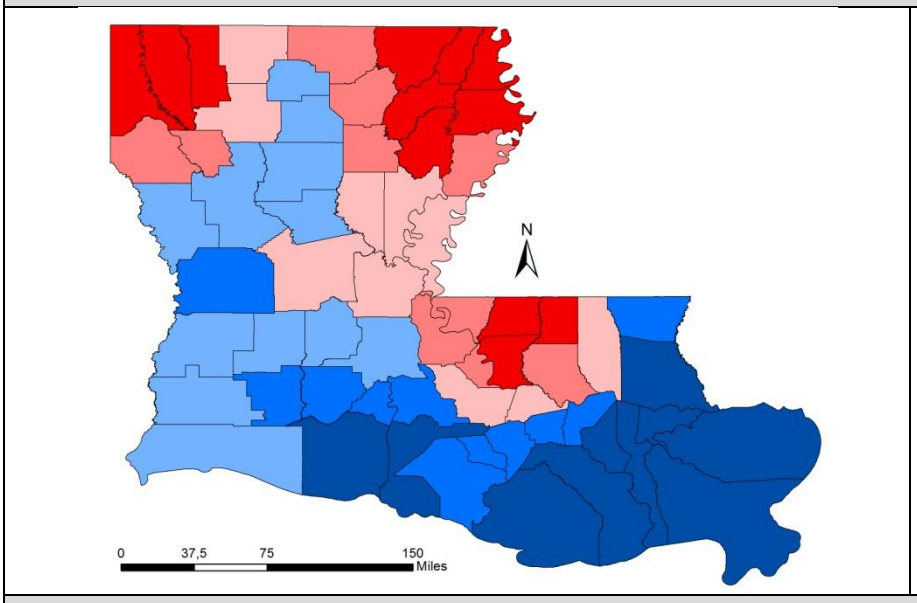


Figure 5.36: Definition of category cutoffs for the legend of the score statistic (taken from Rogerson et al., 2007)

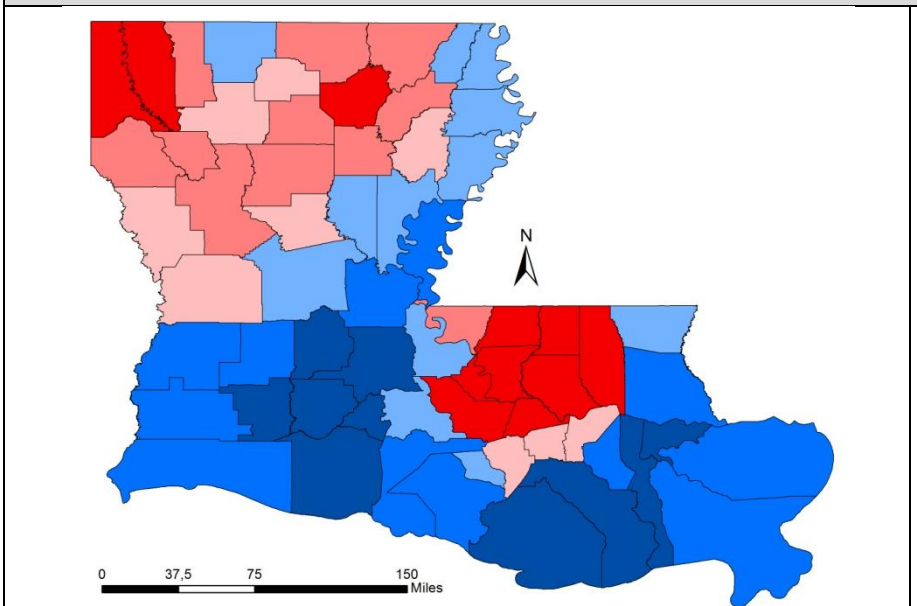




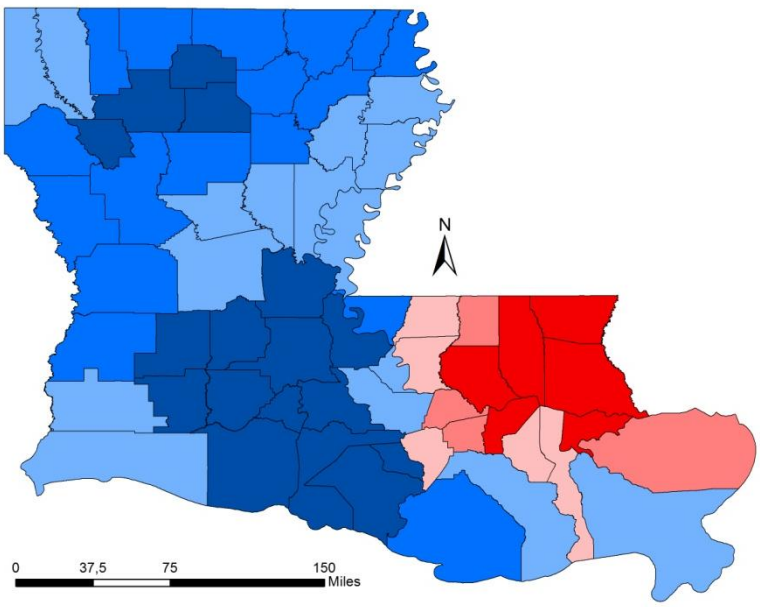
2004



2005



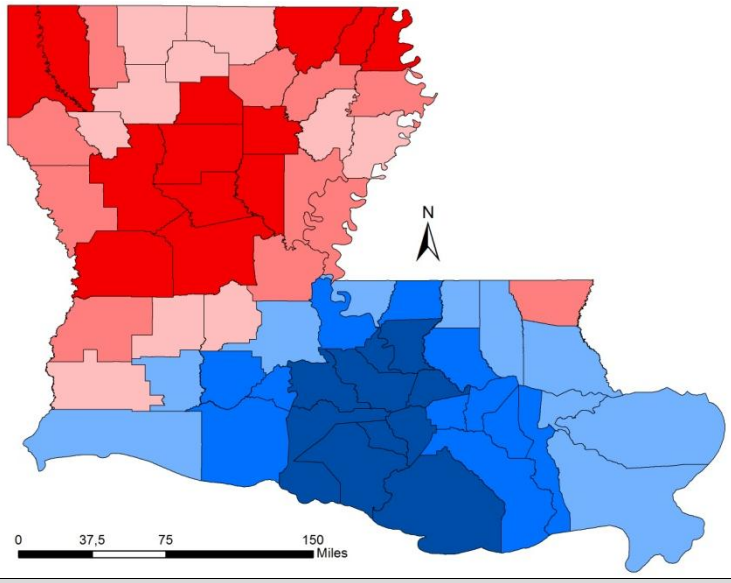
2006



Local score statistic U

-4,89 -	-2,02
-2,01 -	-1,53
-1,52 -	0,00
0,01 -	1,72
1,73 -	3,07
3,08 -	7,83

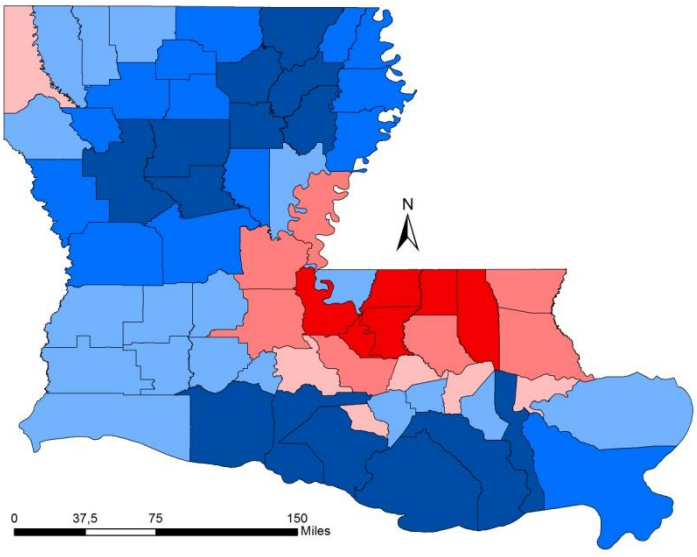
2007



Local score statistic U

-2,42 -	-1,95
-1,94 -	-1,25
-1,24 -	0,00
0,01 -	1,09
1,10 -	1,66
1,67 -	5,07

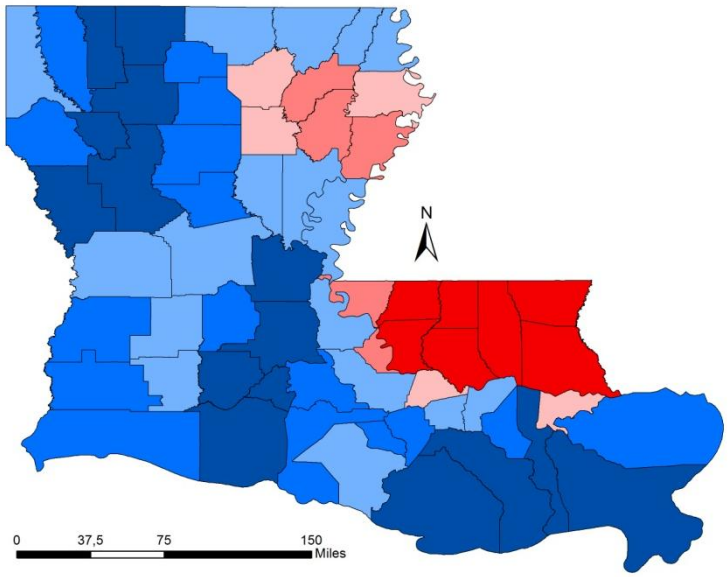
2008



Local score statistic U

-2,19 -	-1,40
-1,39 -	-0,99
-0,98 -	0,00
0,01 -	1,15
1,16 -	2,87
2,88 -	5,86

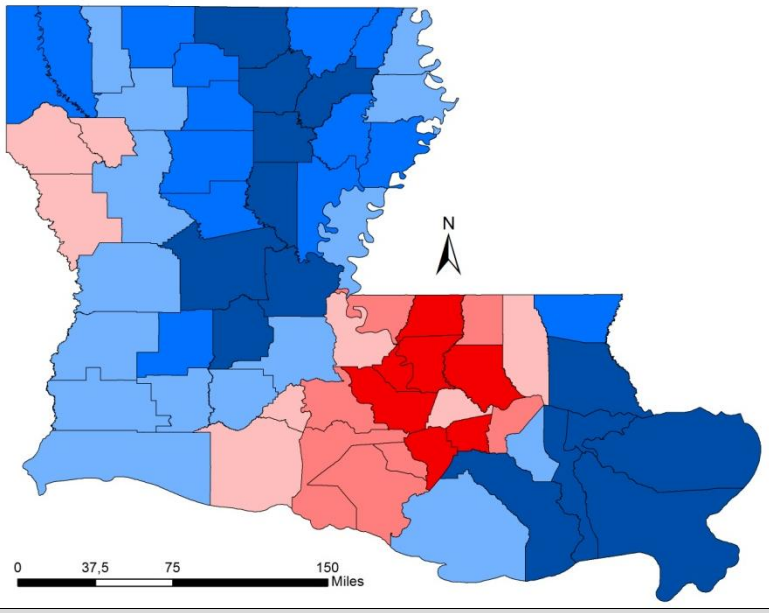
2009



Local score statistic U

-1,76 -	-0,93
-0,92 -	-0,54
-0,53 -	0,00
0,01 -	0,23
0,24 -	1,04
1,05 -	3,82

2010



Local score statistic U

-2,23 -	-1,43
-1,42 -	-0,96
-0,95 -	0,00
0,01 -	1,55
1,56 -	3,62
3,63 -	5,38

2011

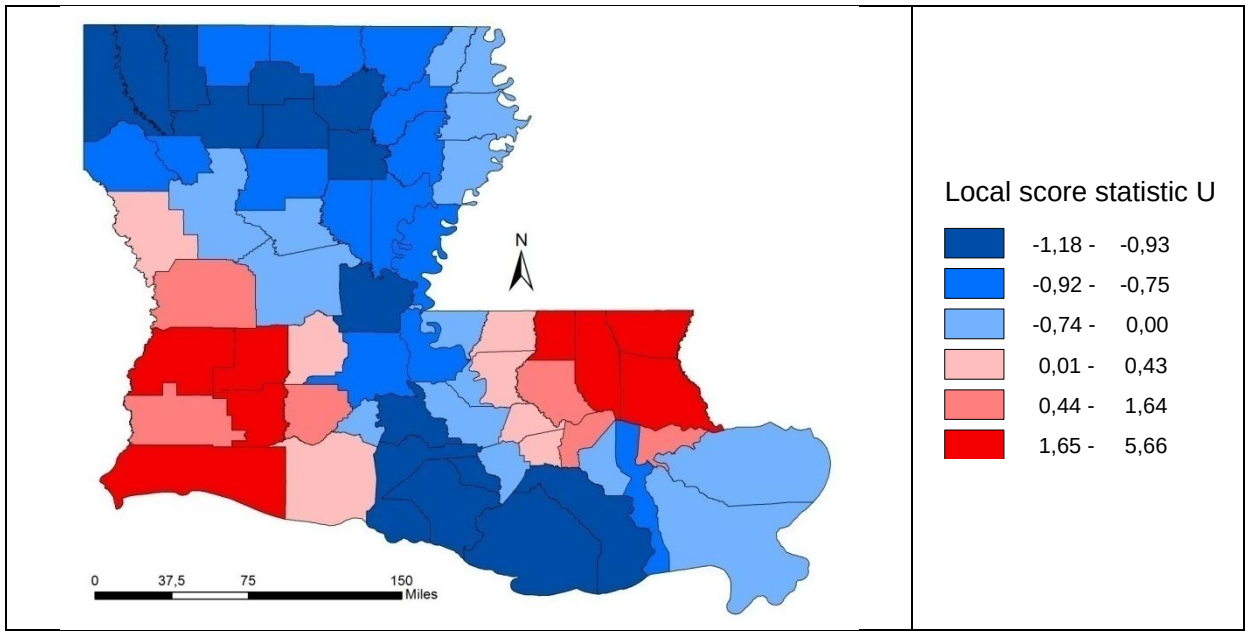


Table 5.3: Retrospective tests of WNV disease data from 2002 till 2011

According to the retrospective analysis in GeoSurveillance, in the year of the WNV outbreak (2002) the parishes with the highest incident rates are located in eastern Louisiana, including the metropolitan area of Baton Rouge. Several parishes form a significant hot spot. The adjacent states are still categorized among the first or even second third of the maximum values. In the northwest there is a large cold spot with values categorized among the minimum second third. In the following arboviral season (2003) patterns shifted. There is still a high incident rate in the eastern part of the state, but the most significant hot spot has evolved in the north-west. Interestingly, one year ago a cold spot has been at the exact same location where the hot spot is now located. In the following years several hot spots emerge and disappear in the north, respectively in the eastern part of Louisiana, suggesting that breeding habitats for both mosquitoes and birds are appropriate enough to maintain the viremia. During those years, cold spots shift between central and south Louisiana. In 2007 WNV incident rates are among the maximum three thirds in the entire northern and central part of Louisiana. After 2007, WNV incidents cease and there are no more hot spots in the northern part of the Gulf state. In contrast, the WNV concentrates in the area, where it initially broke out in 2002, including the eastern parishes and the Baton Rouge metropolitan area. The hot spot distribution pattern between northern and eastern Louisiana faces a severe interruption in the last study year, when unexpectedly a large hot spot emerges in south-western Louisiana. The results of the retrospective test correspond closely with the analysis implemented with Open GeoDa. Especially the hot spots in the eastern and northern part of the state are detected in both analyses. A major difference, however, is noticed in the results for the year 2007. While GeoSurveillance detects a large hot spot that extends across the northern part of Louisiana, Open GeoDa does not find any high-high clusters in that area. Another difference can be seen when comparing both results for the year 2011. While GeoSurveillance classifies the western parishes among the maximum two thirds of the WNV rates, GeoDa suggests that values are low in those parishes but high in their neighbors. These discrepancies might be due to different underlying models, the two software packages are built upon. After all, Open GeoDa is a tool for the detection of spatial autocorrelation in the data, while retrospective testing in GeoSurveillance aims to reject or confirm the null-hypothesis of no spatial clustering.

5.2.2.2 Prospective test

The prospective testing for clusters and clustering is based on the cumulated sum method. The statistic consists of a z-score at a time t and a parameter k. The parameter k is a threshold value, which can be defined in the user interface before calculating the analysis. For the analysis, each total count of a parish in a certain year is converted into a z-value. This is necessary in order to conduct a prospective test. In addition, z-scores have to be ordered in a temporal sequence, thus, each column corresponds to one year. The signal detection for a region is based on both the threshold values k and h. If the value of a z-score exceeds the value of the k-parameter the difference between them is added to the cumulative sum. This step is repeated for each year between 2002 and 2011 and for each parish. If, at the end, the cumulative sum of a parish exceeds the threshold value h the parish is categorized as "Signal". In this analysis the value of the parameters is defined as: k = 1, h = 10. For k, the default value has been adopted. For h, several values have been tested and the chosen value had to be appropriate for the analysis and the used data.

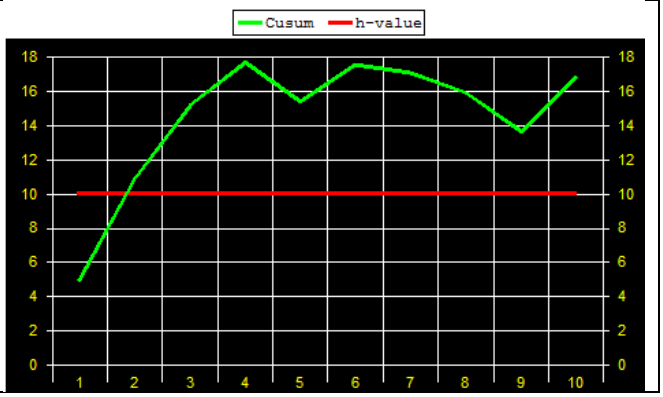
Year	Cusum	Region of maximum cusum value	Chart for maximum cusum x = year, y = cusum value
2002	4.99	Washington	
2003	10.90	Caddo	
2004	15.18	Caddo	
2005	17.73	Caddo	
2006	15.42	Caddo	
2007	17.57	Caddo	
2008	17.09	Caddo	
2009	15.91	Caddo	
2010	13.65	Caddo	
2011	16.83	Washington	

Table 5.4: Maximum cusum value for each year

In Table 5.4 the maximum cusum values for each year are given. In addition, the table depicts the respective parishes where the maximum cusum values have been calculated. Only two parishes are recorded, namely Washington and Caddo. After year 10 (2011), the cusums of five parishes have exceeded the critical threshold value h and are, thus, signal or risk states for prospective WNV outbreaks (see Figure 5.37).

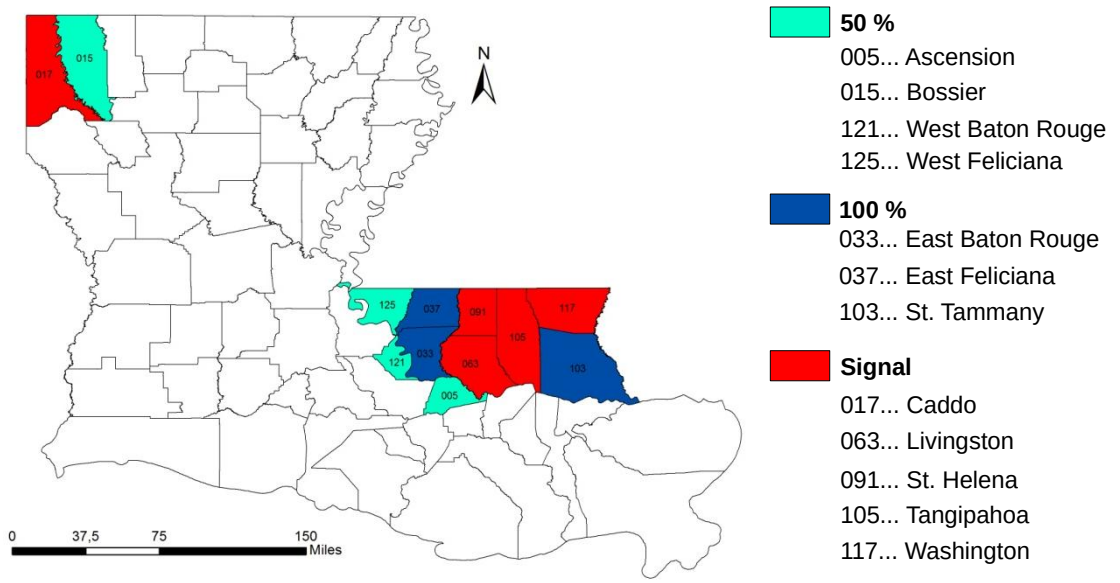


Figure 5.37: Prospective test

The legend for the cusum is constructed based on the threshold (h-value). 25%, 50%, 75% and 100% of the h-value are the class cutoff values used for coloring the map (Rogerson et al., 2007). If the cusum value of a parish exceeds the h-value, it is categorized as “signal”. Apart from Caddo and Washington further signal parishes are Livingston, St. Helena and Tangipahoa. All these regions are located either in the eastern or northern part of the state. This is where hot spots have emerged throughout the last ten years over and over again. East Baton Rouge, East Feliciana and St. Tammany are categorized within 100 % of the h-value. Ascension, Bossier, West Baton Rouge and West Feliciana Parishes are within 50 % of the h-value. Each prospective WNV outbreak location lies either in the east or in the north of the state. Considering these analysis results, precautionary measures and surveillance methods can be implemented in those locations. Prospective tests reveal important information for future predictions in terms of WNV disease distribution patterns.

In addition, GeoSurveillance delivers valuable statistics of the cusum. Thus, the trend of the cusum can be visualized in form of a chart. In table 5.5 the cusum of the five signal states in year 10 (2011) is depicted. The x-axis represents the time from the WNV onset in 2002 till the end of the study period in 2011. The green line is the cusum whereas the red line represents the h-value. The five charts reveal several discrepancies in the evolvement of the WNV. In Caddo Parish there has been a steep increase in disease cases in the first years. The peak is in 2005. Apart from 2008 the cusum starts to gradually decrease. In Livingston Parish, the onset of WNV was not that strong in the first two years. Starting from 2004 the cusum steadily increases, except in 2007, until it reaches an unprecedented peak in 2011. The trend of the line is clearly moving upwards. In St. Helena and Tangipahoa Parishes the trends of the lines are similar to Livingston Parish. Both lines are tending upwards. Washington Parish is the only parish of those five where there was a strong increase in 2002. Afterwards the trend goes downwards till 2005. Apart from that year WNV incidents rise strongly and the trend goes steadily upward till the end of the study period.

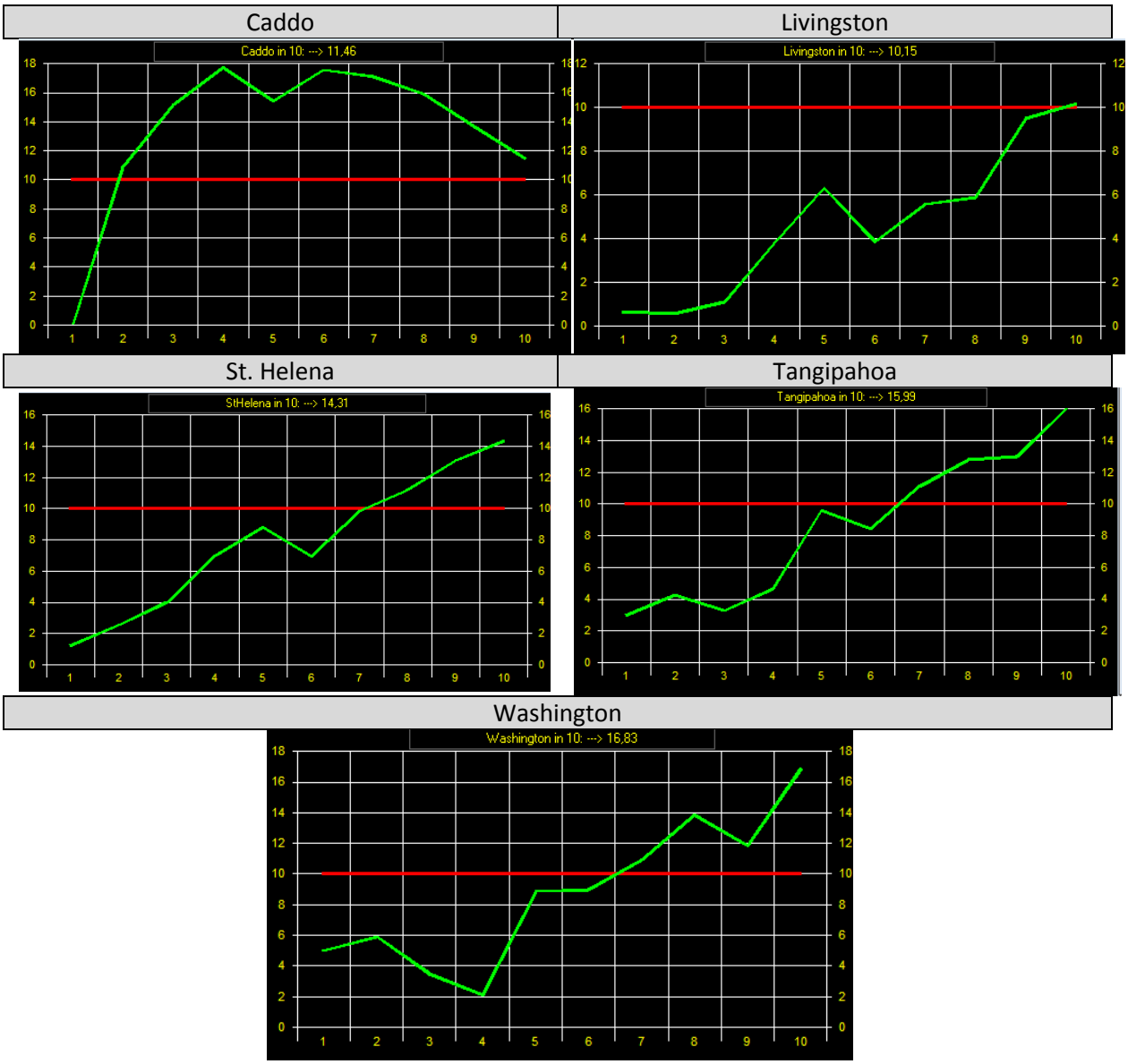


Table 5.5: Cusum for individual regions for the year 10 (2011)

5.2.3 Analysis of space-time clusters of WNV disease data using Kulldorff's Scan Statistic

For the analysis in Louisiana two different probability models have been implemented. Due to their distinct structure the results of the analyses vary significantly. The first probability model is the discrete Poisson model, which is applied to a space-time analysis. It reveals one most likely cluster (cluster 1) in the eastern part of Louisiana, including the Baton Rouge metropolitan area (see Figure 5.38). The lifespan of the cluster starts in 2002 and ends in 2006. In addition, a secondary cluster (cluster 2) is detected in the north-western part of the state, including the Parishes Caddo and Bossier. The secondary cluster has a slightly shorter lifetime (2003 – 2005). In the most likely cluster there are 432 observed cases compared to 144 expected cases. Thus, the ratio between observed and expected is 3 to 1. In the secondary cluster there are 111 observed cases and 26 expected cases. The ratio between observed and expected is 4 to 1.

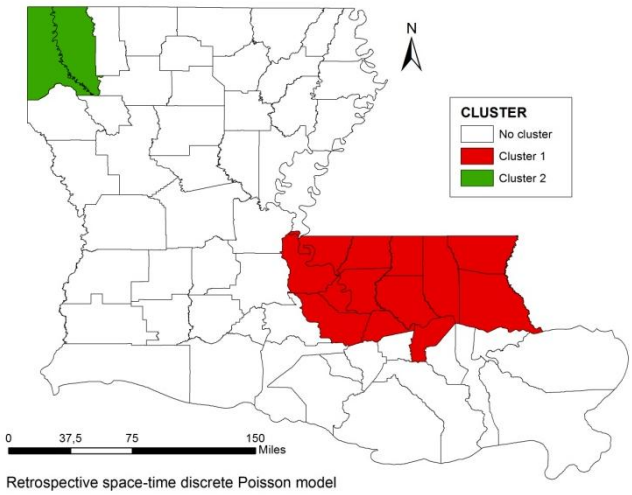


Figure 5.38: Retrospective test (discrete Poisson) for clustering in SaTScan

```
SUMMARY OF DATA

Study period....................: 2002/1/1 to 2011/12/31
Number of locations.............: 64
Total population................: 4455483
Total number of cases...........: 1131
Annual cases / 100000...........: 2.5

MOST LIKELY CLUSTER

1.Location IDs included.: 91, 37, 105, 63, 33, 117, 125, 121, 5,
                          95, 103, 77, 47
  Coordinates / radius..: (30.821975 N, 90.710320 W) / 87.53 km
  Time frame............: 2002/1/1 to 2006/12/31
  Population............: 1184470
  Number of cases.......: 432
  Expected cases........: 144.42
  Annual cases / 100000.: 7.6
  Observed / expected...: 2.99
  Relative risk.........: 4.22
  Log likelihood ratio..: 232.462354
  P-value...............: < 0.000000000000000010

SECONDARY CLUSTERS

2.Location IDs included.: 17, 15
  Coordinates / radius..: (32.580177 N, 93.882359 W) / 28.18 km
  Time frame............: 2003/1/1 to 2005/12/31
  Population............: 360520
  Number of cases.......: 111
  Expected cases........: 26.94
  Annual cases / 100000.: 10.5
  Observed / expected...: 4.12
  Relative risk.........: 4.46
  Log likelihood ratio..: 76.392321
  P-value...............: < 0.000000000000000010
```

Figure 5.39: Summary of space-time analysis (discrete Poisson model) including results for the most likely and secondary clusters

A second retrospective space-time analysis was carried out in SaTScan using the space-time permutation model. This method detects three clusters which have evolved throughout the study period (see Figure 5.40). The most likely cluster (cluster 1) is located in northern Louisiana and extends across several parishes. This is a first major difference to the results of the discrete Poisson model, where the most likely cluster is located in the east. The cluster arises in 2003 and exists till 2005. There are a total of 168 observed cases in that area compared to approximately 92 expected cases, which leads to a ratio 2 to 1. There are two secondary clusters (clusters 2 & 3). Cluster 2 extends from central to southern Louisiana. The cluster includes the Baton Rouge metropolitan area, which is, considering previous analysis results, likely to be a WNV hot spot. However, what makes that secondary cluster dubious is the fact that it includes southern Louisiana parishes, where the WNV has not really been widespread throughout the study period. The result reveals that this cluster has only existed for one arboviral season, which was during the outbreak year in 2002. In that period a total of 30 cases were observed compared to only 8 expected ones (ratio 3.5 to 1). This cluster might have been categorized as significant due to the high observed/expected ratio. Cluster 3 is another secondary cluster, confined to the year 2002. It is located in southeastern Louisiana, including Jefferson Parish, the area where the WNV was initially isolated in 2001. The ratio between observed and expected is only 1.45 (133 versus 92) (see Figure 5.41).

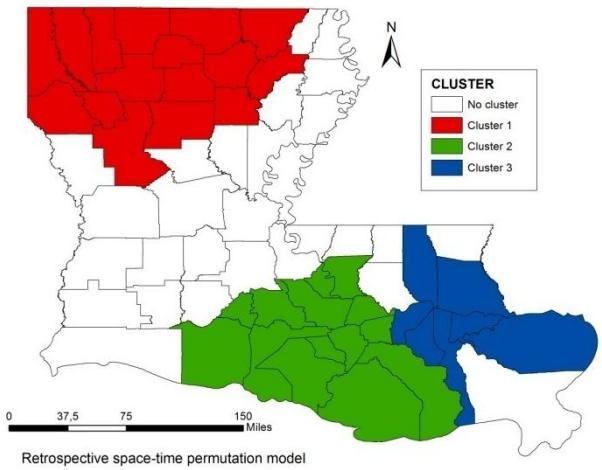


Figure 5.40: Retrospective test for clustering (space-time permutation) in SaTScan

```
SUMMARY OF DATA

Study period....................: 2002/1/1 to 2011/12/31
Number of locations.............: 64
Total number of cases..........: 1131

_____________________________________________________________

MOST LIKELY CLUSTER

1.Location IDs included.: 27, 119, 61, 13, 111, 15, 49, 17, 81,
                          73, 127, 31, 67, 21, 69, 83
  Coordinates / radius..: (32.822693 N, 92.995733 W) / 123.80 km
  Time frame.............: 2003/1/1 to 2005/12/31
  Number of cases.......: 168
  Expected cases........: 92.25
  Observed / expected...: 1.82
  Test statistic........: 27.794233
  P-value...............: 0.00000000000023
```

```
SECONDARY CLUSTERS

2.Location IDs included.: 101, 45, 7, 99, 109, 47, 93, 113, 5,
                          55, 121, 57, 33
  Coordinates / radius..: (29.634623 N, 91.472926 W) / 106.77 km
  Time frame.............: 2010/1/1 to 2010/12/31
  Number of cases.......: 30
  Expected cases........: 8.39
  Observed / expected...: 3.58
  Test statistic........: 16.831139
  P-value...............: 0.00000018

3.Location IDs included.: 71, 103, 51, 89, 95, 87, 105
  Coordinates / radius..: (30.068800 N, 89.930881 W) / 77.77 km
  Time frame.............: 2002/1/1 to 2002/12/31
  Number of cases.......: 133
  Expected cases........: 91.94
  Observed / expected...: 1.45
  Test statistic........: 8.866697
  P-value...............: 0.0035
```

Figure 5.41: Summary of space-time analysis (Space-time permutation model) including results for the most likely and secondary clusters

Similar to the results of retrospective tests for the contiguous U.S.A. the tests for Louisiana do not reveal any “alive” clusters, but only historic clusters. Thus, the prevalence of the WNV is ceasing in this Gulf state. This might be due to increased awareness and precautionary measures or due to a natural reduction of the WNV in that area. For future predictions a prospective analysis was carried out using the prospective space-time discrete Poisson model (see Figure 5.42). The prospective test reveals only one most likely cluster, namely Point Coupee Parish. Interestingly Point Coupee Parish does not lie in any of the considered “hot spot locations” which have been detected in previous analyses. The cluster has emerged in 2008 and persists till the end of the study period. Six cases have been observed, while only 2.3 cases have been expected in that area. That leads to a ratio of 2.3 to 1. The relative risk of being infected with the WNV is 2.6 times higher in the cluster than outside of the cluster. However, the p-value of the cluster is 1.0. Thus, when comparing the p-value to the significance value $\alpha = 0.05$, the p-value is higher than α. Hence, the cluster is not significant (see Figure 5.43).

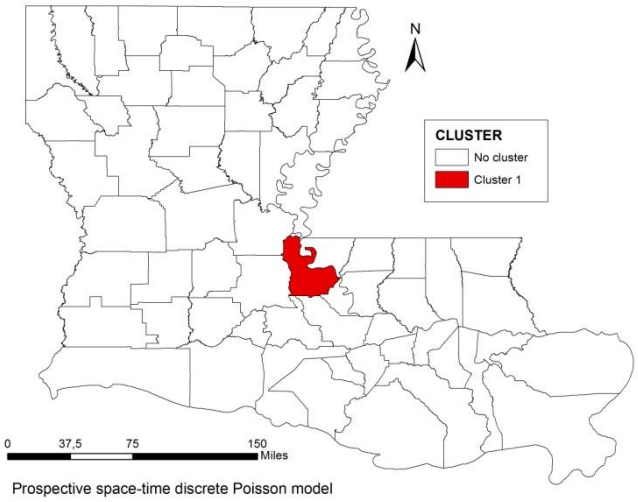


Figure 5.42: Prospective space-time clustering in SaTScan

```
MOST LIKELY CLUSTER

1.Location IDs included.: 77
  Coordinates / radius..: (30.709396 N, 91.600789 W) / 0 km
  Time frame............: 2008/1/1 to 2011/12/31
  Population............: 22389
  Number of cases.......: 6
  Expected cases........: 2.30
  Annual cases / 100000.: 6.6
  Observed / expected...: 2.61
  Relative risk.........: 2.62
  Log likelihood ratio..: 2.059888
  P-value...............: 1.00
  Recurrence interval...: 1.0 year
```

Figure 5.43: Summary of the most likely (but not significant) prospective cluster in Louisiana

5.2.4 Visualization of univariate and multivariate space-time patterns in Vis-Stamp

The Louisiana Department of Health and Hospitals provides data about human WNV disease cases, both fever and neuroinvasive diseases. In addition, their annual reports include information about dead avian species, sentinel chicken, equine disease cases and infected mosquito pools. This data, however, is not completely available for the entire study period. As far as sentinel chicken is concerned, its collection and testing methods have been declared inefficient by the department (DHH, 2012). Thus, DHH stopped testing chicken for WNV infections. Mosquito pools are not tested on a regular basis. The testing intervals also vary from parish to parish, depending on financial resources (DHH, 2012). Thus, “infected mosquito pools” would bias analysis results. There is no way to normalize it, thus, was excluded from this analysis. The remaining variables (dead avian species, sentinel chicken and equine disease cases) are completely available from 2002 till 2005. For this period, a multivariate analysis can be carried out.

First, univariate space-time patterns are identified. This analysis is based on the total cases of WNV incidents in a parish compared to its population. For this purpose a SOM with the dimensions 7x7 is created. The SOM assigns similar clusters similar colors (see Figure 5.44). The SOM depicts a large blue circle in the upper right corner. This is the largest cluster in terms of number of parishes. This circle represents clusters with little WNV activity. In the opposite corner, the significantly smaller red circle represents clusters with the highest number of WNV incidents in the study area and study period. The fact that the blue circle appears in a white hexagon while the red circle is located within a dark shaded hexagon is an indicator for large dissimilarities between two clusters.

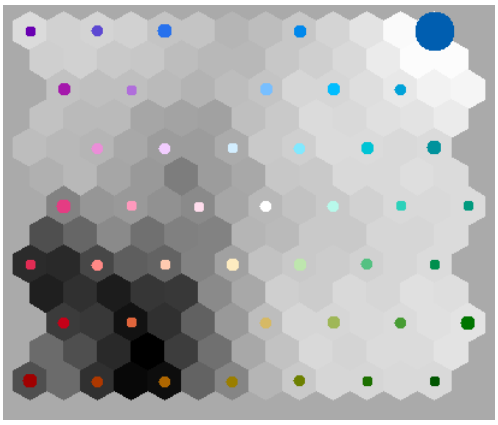


Figure 5.44: 7x7 SOM, coloring clusters of spatial objects

In the space-time matrix patterns over space and time are depicted (see Figure 5.45). The matrix is built with the years of the study period (columns) and the parishes (rows). Vis-Stamp orders the parishes in a way, that similar patterns lie close to each other. This allows quick cluster detection. For example, parishes with a high onset of WNV cases in the outbreak year and with continually high WNV rates in the following years are listed in the upper part of the matrix. Thus, a cluster has evolved in the period from 2002 till 2006 in the parishes Point Coupee, Tangipahoa, St. Tammany, Washington, Ascension, Livingston and St. Helena. A second large cluster which evolved in 2004 and persisted until 2005 can be seen in the lower part of the matrix. This cluster involves the parishes Franklin, Caldwell, Natchitoches, Ouachita, West Baton Rouge, East Baton Rouge, Rapides and Iberville. Until 2008 several smaller clusters appear which exist only for one period. There is a cluster in 2002 involving ten parishes. Weaker clusters with low WNV incidents and including just a few parishes emerge in 2003, 2006, and in 2008 (Point Coupee, Tangipahoa, St. Tammany).

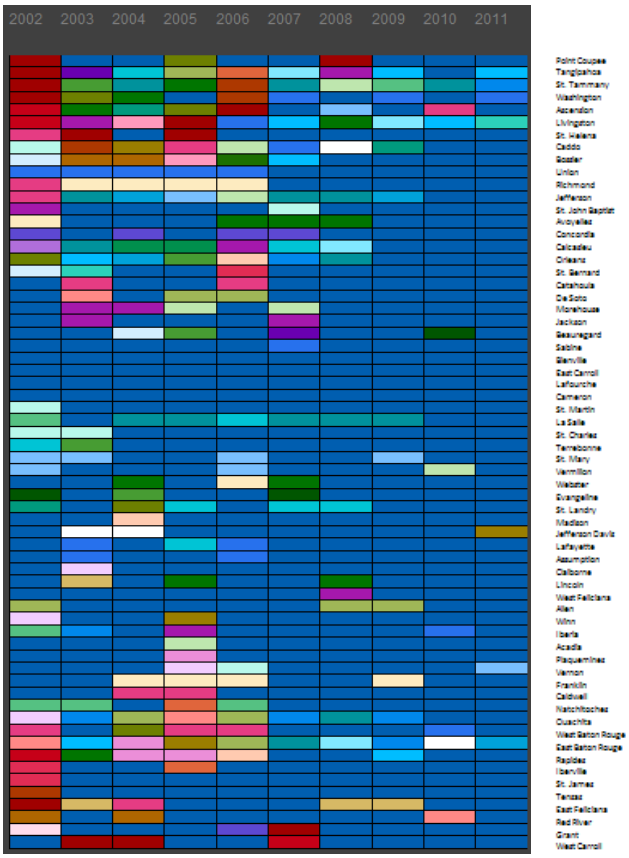


Figure 5.45: Space-time matrix

The PCP highlights the clusters and assigns them values (see Figure 5.46). This analysis is univariate. The variable investigated is the rate of WNV cases per 10,000 population. The cluster with no cases is represented as a thick blue line, which indicating a high number of members in that cluster. There is a cluster with 1.5 cases per 10,000 population, colored in bright red. Also this cluster has a high number of parishes.

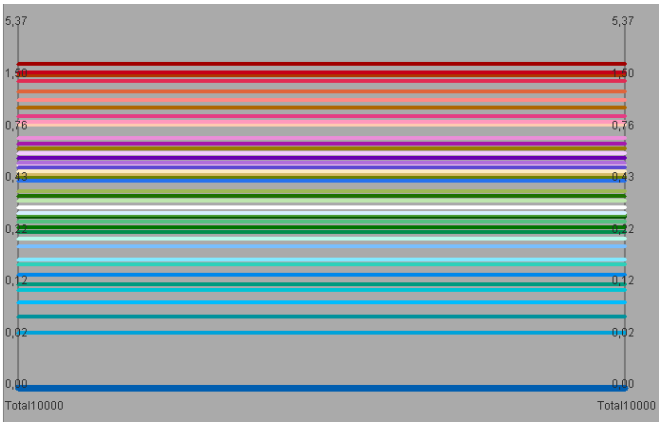


Figure 5.46: PCP

The map matrix displays spatial patterns year by year. This makes interpretation easier. In 2002 there is a large hot spot in eastern Louisiana. Several clusters with slightly elevated WNV incident rates are scattered all over the state, excluding the very northern and western parishes. In the following years clusters with less significantly and significantly elevated cases are scattered from eastern to northern Louisiana. Especially the eastern Parishes St. Tammany, Washington, Tangipahoa, East and West Baton Rouge are affected significantly in

the years 2005 and 2006. After 2008 there is a strong reduction in WNV incidents. Thereafter, the WNV keeps ceasing in the state.

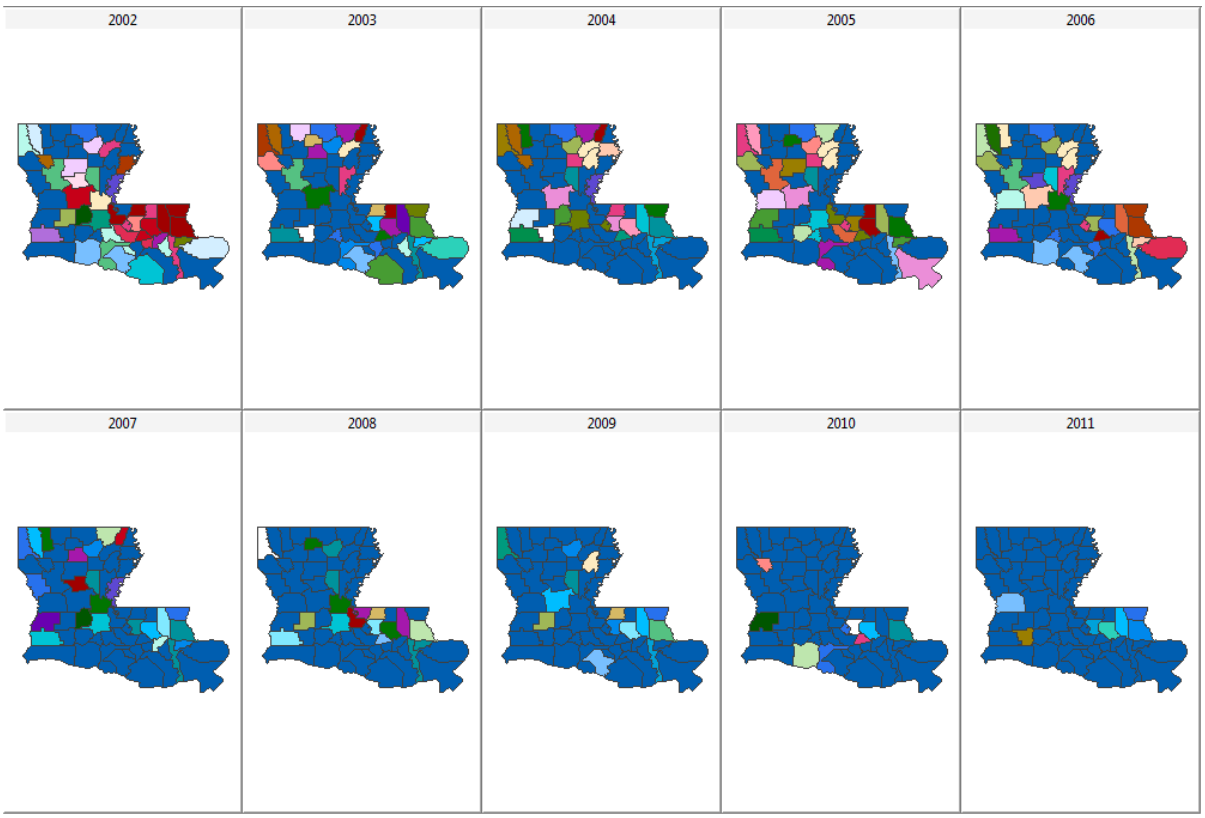


Figure 5.47: Map Matrix

Since the WNV is generally associated with avian die-offs and disease cases in domestic animals, the following multivariate analysis is carried out in order to detect possible connections between human and animal cases. Apart from human WNV disease rates per 10,000 population, the variables “bird”, “sentinel chicken” and “horse” are included in the analysis. Incidents among these species have been collected and reported by the Louisiana Department of Health and Hospitals (DHH, 2012).

The higher number of variables makes the PCP look slightly different (see Figure 5.48). The x-axis of the plot depicts the variables while the y-axis is divided into cluster category cutoff values. Each cluster is represented by a colored line, which has certain attributes. For example, if a parish is colored in bright red it has at least ten sentinel chicken occurences, 29 dead birds, the WNV has been diagnosed in more than five horses and there have been at least three human WNV disease cases among 10,000 people. This cluster suggests that there is a high connection between those four variables. The white cluster represents a high number of cases among birds and several human cases. The pink and purple cluster stand for a low human incident rate but high rates among the other species, while the green clusters represent the opposite (high human incident rate and low rates among chicken, birds and horses). Blue clusters, in general, represent low rates across all variables.

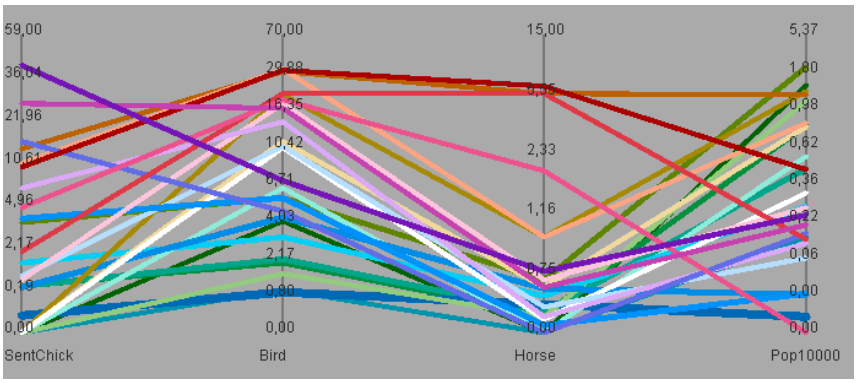


Figure 5.48: Multivariate PCP

The space-time matrix reveals a large green cluster in the outbreak year (see Figure 5.49). The parishes belonging to that cluster have a high human WNV disease incident rate. However, animal cases are scarce. A second cluster emerges in that same year (2002). It is a purple cluster, representing very high sentinel chicken cases, several bird cases, and little human and equine cases. This cluster persists till the next arboviral season. In 2003, a white, green, and orange pattern emerges. These clusters have one thing in common, namely very high human disease rates. The most affected locations are Bossier, Caddo, West Carroll and St. Helena. The cluster type changes in the following year as human cases decrease. In 2005, there is another green-red-orange pattern, including Iberville, Ouachita, Livingston, and St. Helena.

If there is a connection between animal and human incident cases, clusters would either be red or orange representing high cases in every variable or they would be blue which means low cases across all parameters. However, this is not the case, because there are plenty of white, purple and green clusters. One reason for this is that the WNV does not spread uniformly because it depends on vectors and hosts. Another issue that has to be put into consideration is that the location where a mosquito contracts the WNV to a bird and the location where the bird dies due to WNV might be completely different since many avian species migrate. Also, there might be a large time span between these two events, consisting of the incubation period and the development of high enough viremia which is eventually lethal for the bird. A pattern which suggests a connection between human and animal cases would be, first a purple or pink cluster (high cases among chicken and birds) directly followed by a green, orange red or white cluster (high human cases).

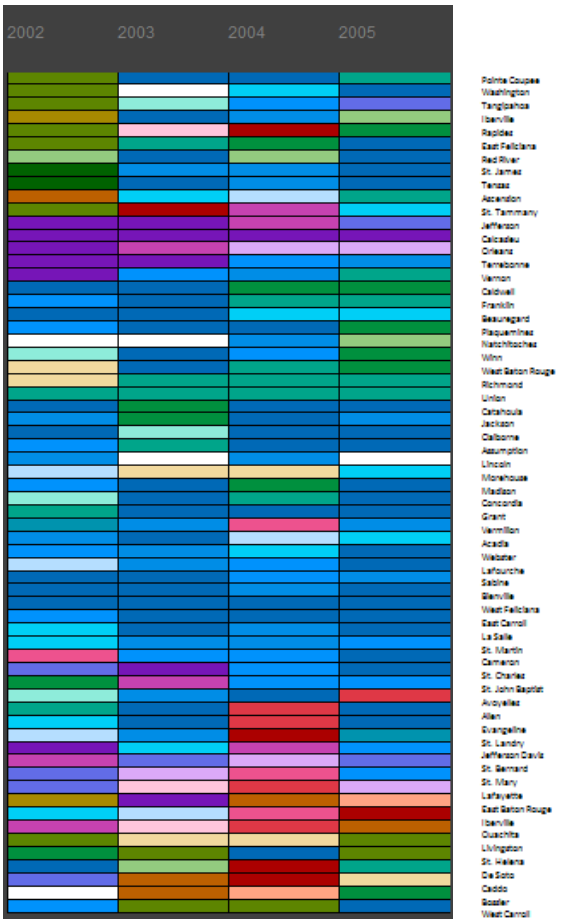


Figure 5.49: Space-time matrix

The space-time patterns are visualized in the form of maps in the map matrix (see Figure 5.50). Interestingly, in the outbreak year in 2002 a large green cluster with high human disease rates but low animal incident cases evolved.

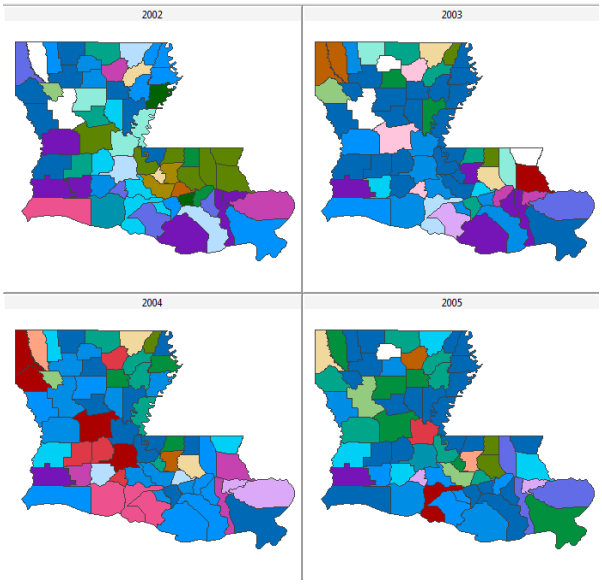


Figure 5.50: Map matrix

6 Conclusion

The WNV has been a severe danger for both humans and animals over more than two decades. Since its initial introduction into the U.S. in New York City in 1999 it has spread across the entire contiguous U.S. Cases have also been reported in Canada apart from 2002 (PHAC, 2011) and in the Caribbean apart from 2007 (Barrera et al., 2008). The successful spread of the WNV depends on several circumstances. First, areas with optimal breeding sites for birds and mosquitoes provide a decent environment for the WNV to evolve. A second circumstance is the prevalence of a sentinel population at risk. Humans and mammals are so-called incidental hosts in the natural transmission cycle of the virus. If there is an outbreak of the WNV in an unpopulated area the virus does not form any harm to humans and domestic animals unless it moves on to more populated areas. However, the virus is considered an urban pathogen (Ruiz et al., 2007), endangering avian species and susceptible to human risk groups like elderly people. The introduction of the WNV into the U.S. is still a controversial issue. Most theories state that the ease of intercontinental travel methods is primarily responsible for the spread of the WNV to the western hemisphere (Sfakianos, 2005). While the spread of the WNV inside the U.S. succeeded in a slow way from 1999-2001, the spread of the WNV suddenly exploded starting in 2002. In that year same, both the number of affected regions as well as the number of affected people diagnosed with the WNV disease rose to an unprecedented level. In the following season (2003) incident cases even doubled. The reason for this sudden outbreak might be associated with the climate. In general, climate is the third circumstance for the evolvement and success of a virus. The winter season of 2001 and 2002 was quite warm. Precipitation in the form of snow levels was below the average. This led to water shortages in the western U.S. due to a lack of spring melt water (Weather explained, 2012). The 2002 summer season was also one of the warmest on record. The year 2002 has been considered to be the second warmest year since 1881 (the beginning of the records). This might be due to a strengthening El Niño episode, in late boreal summer continuing into early winter (Weather explained, 2012). The mild winter and the high temperatures as well as water shortages in spring might have had a positive effect on mosquito reproductions. Thus, many WNV infected species could have been able to overwinter, carrying the virus into a new season under better circumstances. In addition, there could have been an increase of breeding sites for mosquitoes due to the high temperatures and scattered drought episodes. For instance, the culex mosquito prefers to breed in small bodies of water (Weather explained, 2012). The analyses in this thesis show that the WNV spreads from the U.S. east to the U.S. north coast within a matter of three years, hitting California severely in 2004. During the course of this rapid spread, several hot spots of WNV human disease cases evolved. In the near future, prospective clusters might emerge in southern states like Louisiana, Mississippi, Arizona, and New Mexico. However, also the state of Idaho is considered to be part of a most likely cluster for a WNV disease outbreak in the near future. The analyses of human WNV cases also reveal that in the recent five years the prevalence and the severity of the virus has ceased steadily. This observed trend corresponds to the nature of the virus which usually has a strong initial emergence before slowly ceasing again. The cease can also be due to better precautionary measures, improved surveillance systems, and increased public awareness. In Louisiana, like in many other states, the WNV broke out in 2002. The first case was found in a homeless man in 2001 in Jefferson Parish. Since then, the WNV spread across the entire state forming hot spots in northern and south-eastern Louisiana. In Louisiana, the Department of Health and Hospitals is responsible for the surveillance of the WNV. They have issued annual reports about WNV

activity in the state. Apart from human disease cases also bird deaths, WNV-infected chicken, and horses are recorded. There is no significant correlation between human cases and WNV sentinel chicken cases, or bird deaths, respectively. This might be due to the fact that many avian species migrate. Studying the routes of avian species is a future project outlined by the CDC (Rappole et al., 2000). Thus, scientists can investigate transmission routes created by avian species, which are most likely responsible for transmitting the WNV to the Caribbean and even to parts of Latin America.

7 References

7.1 Literature

Eck, J., Chainey, S., Cameron, J., Leitner, M., Wilson, R. (2005). Mapping Crime: Understanding Hot Spots, U.S. Department of Justice, National Institute of Justice – Special Report, August 2005

ESRI (2011). ArcGIS Desktop: Release 10. Redlands, CA: Environmental Systems Research Institute

Gan, G., Ma, C., Wu, J. (2007). Data Clustering, Theory, Algorithms and Applications, ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, VA, 2007

Gruszynski, K. (2006). The epidemiology of West Nile Virus in Louisiana, A Dissertation submitted to the graduate faculty of the Louisiana State University, Louisiana State University Health Sciences Center, December 2006

Levine, N. (2004). CrimeStat III, A spatial statistics program for the analysis of crime incident locations. Ned Levine & Associates, Houston, TX, and the National Institute of Justice, Washington D.C., November 2004

Modrow, S., Falke, D., Truyen, U. (2003). Molekulare Virologie, 2. Auflage, Spektrum, Heidelberg 2003

Scheld, M., Hooper, D., Hughes, J. (2007). Emerging Infections 7, 2007 ASM Press, American Society for Microbiology, Washington D.C.

Sfakianos, J.(2005). West Nile Virus, Deadly Diseases and Epidemics, Chelsea House Publishers, New York, 2005

White, D.; Morse, D. (2001). West Nile Virus, Detection, Surveillance and Control, Annals of the New York Academy of Science, Volume 952, December 2001

Xu, R., Wunsch D. (2009). Clustering, IEEE Computational Intelligence Society, IEEE Press Series on Computational Intelligence, Hoboken, New Jersey, 2009

7.2 Online Resources

Anselin, L., Cohen, J., Cook, D., Gorr, W., Tita, G. (2000). Spatial analyses of crime. In: David Duffee (ed.), Criminal Justice 2000, Measurement and analysis of crime and justice, Voume 4, National Institute of Justice, Washington D.C., 2000; Link: http://www.geo.hunter.cuny.edu/~dougwill/CRIME/web_pdfs/spatial_analysis_of_Crime-anselin.pdf (last accessed on April 20, 2012)

Anselin, L. (2003a). Review of Cluster Analysis Software, Anselin and Associates, LLC., Urbana, IL, August 2003; Link: http://www.schs.state.nc.us/NAACCR-GIS/pdfs/clustersoftwareFinal.pdf (last accessed on April 21, 2012)

Anselin, L. (2003b). GeoDa 0.9 User's Guide, Spatial Analysis Laboratory, Department of Agricultural and Consumer Economics, University of Illinois, Urbana, IL, June 2003; Link: http://www.unc.edu/~emch/gisph/geoda093.pdf (last accessed on May 5, 2012)

Anselin, L., (2003c). An Introduction to EDA with GeoDa, Department of Agricultural and Consumer Economics, University of Illinois, Urbana, IL, June 2003; Link: http://www.spatial.maine.edu/~beard/quicktour.pdf (last accessed on May 7, 2012)

Anselin, L., Ibnu, S., and Youngihn, K. (2004). GeoDa: An Introduction to Spatial Data Analysis. Geographical Analysis; Link: http://geodacenter.asu.edu/pdf/geodaGA.pdf (last accessed on May 5, 2012)

Barrera, R., Hunsperger, E., Muñoz-Jordán, J., Amador, M., Diaz, A., Smith, J., Bessoff, K., Beltran, M., Vergne, E., Verduin, M., Lambert, A., Sun, W. (2008). First isolation of West Nile Virus in the Caribbean, Am J Trop Med Hyg., April 2008; Link: http://www.ncbi.nlm.nih.gov/pubmed/18385366 (last accessed on June 2, 2012)

BioMedware (2012a). Official Website of BioMedware, Geospatial research and software, ClusterSeer, Polygon contiguity, Link: http://www.biomedware.com/files/documentation/clusterseer/Concepts/Polygon_contiguity.htm (last accessed on May 5, 2012)

BioMedware (2012b). Official Website, Geospatial research and software, SpaceStat, Spatial Weights Details, Link: http://www.biomedware.com/files/documentation/spacestat/data/adj/Spatial_Weights_Details.htm (last accessed on May 5, 2012)

Bowden, S., Magori, K., Drake, J. (2011). Regional differences in the association between land cover and West Nile Virus disease incidence in humans in the United States; The American Society of Tropical Medicine and Hygiene, Am. J. Trop. Med. Hyg., 84(2), 2011; Link: http://www.ajtmh.org/content/84/2/234.full (last accessed on April 12, 2012)

CDC (2009). Official Website of the Centers for Disease Control and Prevention, Vertebrate Ecology; Link: http://www.cdc.gov/ncidod/dvbid/westnile/birds&mammals.htm (last accessed on March 15, 2012)

CDC, (2009a). Center of Disease Control and Prevention, West Nile Virus, Surveillance and Epidemiology in the United States and Tropical Americas, West Nile Virus Conference 2009, Fort Collins, CO; Link: http://www.slideserve.com/osgood/west-nile-virus-surveillance-and-epidemiology-united-states-and-tropical-americas (last accessed on April 4, 2012)

CDC (2011a). Official Website of the Centers for Disease Control and Prevention, Maps & Human Cases; Link: http://www.cdc.gov/ncidod/dvbid/westnile/index.htm (last accessed on June 8, 2012)

CDC (2011b). Official Website of the Centers for Disease Control and Prevention, Arboviral Diseases, Neuroinvasive and Non-Neuroinvasive 2011 Case Definition; Link: http://www.cdc.gov/osels/ph_surveillance/nndss/print/arboviral_current.htm (last accessed on March 14, 2012)

Cooke, W., Grala, K., Wallis, R. (2006). Avian GIS models signal human risk for West Nile Virus in Mississippi, International Journal of Health Geographics 5:36, August 2006; Link: http://www.ncbi.nlm.nih.gov/pubmed/16945154 (last accessed on April 8, 2012)

DHH (2002). West Nile Virus Surveillance Update in Louisiana, 2002 Louisiana Department of Health and Hospitals; Link: http://new.dhh.louisiana.gov/index.cfm/page/539 (last accessed on May 21, 2012)

DHH (2003). West Nile Virus Surveillance Update in Louisiana, 2003 Louisiana Department of Health and Hospitals; Link: http://new.dhh.louisiana.gov/index.cfm/page/539 (last accessed on May 21, 2012)

DHH (2004). West Nile Virus Surveillance Update in Louisiana, 2004 Louisiana Department of Health and Hospitals; Link: http://new.dhh.louisiana.gov/index.cfm/page/539 (last accessed on May 21, 2012)

DHH (2012). Official Website of the Louisiana Department of Health and Hospitals, Center for Community and Preventive Health, West Nile Virus; Link: http://new.dhh.louisiana.gov/index.cfm/page/539 (last accessed May 28, 2012)

Diggle, P. (2003). Statistical analysis of spatial point patterns, second edition, Hodder Education Publishers, New York, February 2003; Link: http://www.cabnr.unr.edu/weisberg/NRES675/Diggle2003.pdf (last accessed on April 14, 2012)

Fuchs, C., Kenett, R. (1980) A test for detecting outlying cells in the multinomial distribution and two-way contingency tables. Journal of the American Statistical Association, 75, 395–8; Link: http://www.springerlink.com/content/g26m83p6u3385k65/ (last accessed on May 7, 2012)

Guo, D., Gahegan, M., MacEachren, A., Zhou, B. (2005). Multivariate analysis and geovisualization with an integrated geographic knowledge discovery approach, NIH public access, Cartogr Geogr Inf Sci. April, 2005; Link: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2786224/ (last accessed on May 5, 2012)

Guo, D. (2006a). Vis-Stamp: A visualization system for space-time and multivariate patterns; Link: http://www.spatialdatamining.org/software/visstamp (last accessed on May 28, 2012)

Guo, D., Chen, J., MacEachren, A., Liao, K. (2006); A visualization system for space-time and multivariate patterns (Vis-Stamp), IEEE Transactions on visualization and computer graphics, Volume 12, Number 6, November/December 2006; Link: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3170656/ (last accessed on April 30, 2012)

Guo, D. (2009). VIS-STAMP: A visualization system for space-time and multivariate patterns, User Manual Version 1.0, Department of Geography, University of South Carolina, June 2009; Link: http://www.spatialdatamining.org/software/visstamp (last accessed on May 21, 2012)

Griffith, D. (2009). Spatial Autocorrelation, University of Texas at Dallas, Richardson, TX, 2009; Link: http://www.elsevierdirect.com/brochures/hugy/SampleContent/Spatial-Autocorrelation.pdf (last accessed on May 5, 2012)

Huhn, G., Sejvar, J., Montgomery, S., Dworkin, M. (2003). West Nile Virus in the United States: An Update on an Emerging Infectious Disease, American Family Physician, Volume 68, Number 4, August 2003; Link: http://www.aafp.org/afp/2003/0815/p653.html (last accessed on March 29, 2012)

Infoplease (2012). Official Website of Information Please, U.S. statistics, population; Link: http://www.infoplease.com/ipa/A0004986.html (last accessed on April 14, 2012)

Kulldorff, M. (2011). SaTScan: Software for the spatial, temporal and space-time scan statistics; Link: http://www.satscan.org/ (last accessed on June 5, 2012)

Levine, N. (2005). CrimeStat III, Documentation, Part II: Spatial description, Chapter 6 – 'Hot Spot' Analysis, March 2005; Link: http://www.icpsr.umich.edu/CrimeStat/download.html (last accessed on April 28, 2012)

Lobigs, M., Larena, M., Alsharifi, M., Lee, E., Pavy, M. (2009). Live Chimeric and Inactivated Japanese Encephalitis Virus vaccines differ in their protective values against Murray Valley Encephalitis Virus, Journal of Virology, Volume 83, Number 6, March 2009; Link: http://jvi.asm.org/content/83/6/2436.full (last accessed on March 28, 2012)

Louisiana.gov (2012a). Official Website of the State of Louisiana, About Louisiana; Link: http://louisiana.gov/Explore/About_Louisiana/ (last accessed on April 17, 2012)

Louisiana.gov (2012b). Official Website of the State of Louisiana, Demographics and Geography; Link: http://louisiana.gov/Explore/Demographics_and_Geography/ (last accessed on April 17, 2012)

Marr, J., Calisher, C. (2003). Alexander the Great and West Nile Virus Encephalitis, Emerging Infectious Diseases, Volume 9, Number 12, December 2003; Link: http://wwwnc.cdc.gov/eid/article/9/12/03-0288_article.htm (last accessed on April 10, 2012)

Mosquito-Netting (2012). Official Website of the Mosquito Netting Project, All Mosquito Netting Info; Link: http://www.mosquito-netting.com/mosquito-larvae.html (last accessed on March 18, 2012)

Kohonen, T. (1990). The self-organizing map, Proceedings of the IEEE, Volume 78, Number 9, September 1990; Link: http://www.eicstes.org/EICSTES_PDF/PAPERS/The%20Self-Organizing%20Map%20(Kohonen).pdf (last accessed on May 17, 2012)

Kulldorff, M. (1997). A spatial scan statistic, Biometry Branch, DCPC, National Cancer Institute, Bethesda, MD, December 1997; Link: http://www.satscan.org/papers/k-cstm1997.pdf (last accessed on May 15, 2012)

Kulldorff, M. (2010). SaTScan User Guide for version 9.0, Department of Ambulatory Care and Prevention, Harvard Medical School, Boston, MA, July 2010; Link: http://www.satscan.org/techdoc.html (last accessed on May 15, 2012)

Neill, D., Moore, A., Sabhnani, M., Daniel, K. (2005). Detection of emerging space-time clusters, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, August 2005; Link: http://www.cs.cmu.edu/~neill/papers/sss-kdd05.pdf (last accessed on April 19, 2012)

Petersen, L., Marfin, A. (2002). West Nile Virus: A Primer for the Clinician, Annals of Internal Medicine, Volume 137, Number 3, August 2002; Link: http://www.annals.org/content/137/3/173.full (last accessed on April 2, 2012)

Petersen, L., Roehrig, J., Sejvar, J.(2009). West Nile Encephalitis Virus Infection, Chapter 1: Global Epidemiology of West Nile Virus; Link: http://www.springerlink.com/content/u485813712q6616t/ (last accessed on April 18, 2012)

PHAC (2011). Official Website of the Public Health Agency of Canada, Maps & Stats, West Nile virus MONITOR; Link: http://www.phac-aspc.gc.ca/wnv-vwn/index-eng.php (last accessed on June 2, 2012)

Rappole, J., Derrickson, S., Hubálek, Z. (2000). Migratory birds and spread of West Nile Virus in the western hemisphere, Emerg Infect Dis Journal, Volume 6, Number 4, August 2000; Link: http://wwwnc.cdc.gov/eid/article/6/4/00-0401_article.htm#suggestedcitation (last accessed on June 2, 2012)

Rochlin, I., Turbow, D., Gomez, F., Ninivaggi, D., Campbell, S. (2011). Predictive Mapping of human risk for West Nile Virus (WNV) based on environmental and socioeconomic factors, PLoS ONE, Volume 6, Issue 8, August 2011; Link: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3154328/pdf/pone.0023280.pdf (last accessed on April 8, 2012)

Rogerson, P. (2005). A set of associated statistical tests for spatial clustering, Environmental and Ecological Statistics, Volume 12, Number 3, Buffalo, NY, September 2005; Link: http://www.springerlink.com/content/g26m83p6u3385k65/ (last accessed on May 7, 2012)

Rogerson, P., Lee, G., Yamada, I. (2007). GeoSurveillance 1.1 User's Manual, National Center for Geographic Information and Analysis, Department of Geography, State University of New York at Buffalo, December 2006; Link: www.acsu.buffalo.edu (last accessed on May 11, 2012)

Rogerson, P., Yamada, I., and Lee, G. (2009). *GeoSurveillance*: a GIS-based Monitoring System for Detection of Spatial Clusters, Journal Geogr Syst 2009 11:155-173; Link: http://xa.yimg.com/kq/groups/13354653/1183984305/name/geosurvillance.pdf (last accessed on May 25, 2012)

Ruiz, M., Walker, E., Foster, E., Haramis, L., Kitron, U. (2007). Association of West Nile Virus illness and urban landscapes in Chicago and Detroit, International Journal of Health Geographics 6:10, March 2007; Link: http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=Link&db=pubmed&dbFrom=PubMed&from_uid=17352825&holding=f1000%2Cf1000m%2Cisrctn (last accessed on April 12, 2012)

The New York Times (2010). The New York Times Health Guide, West Nile Virus; Link: http://health.nytimes.com/health/guides/disease/west-nile-virus/overview.html#Symptoms (last accessed on April 4, 2012)

Taylor Enterprises, Inc. (2012). Official Website, Cumulative sum chart (Cusum); Link: http://www.variation.com/cpa/help/hs108.htm (last accessed on May 11, 2012)

Tiny Mosquito (2012). Official Website for information about mosquitoes and larvae; Dangers of the Culex Mosquito; Link: http://www.tinymosquito.com/culex.html (last accessed on March 18, 2012)

Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region, Economic Geography 46:234–40; Link: http://www.geographdy.com/blog/wp-content/uploads/2009/09/sui_2004.pdf (last accessed on May 5, 2012)

U.S. Census (2010). Official Website of the U.S. Census Bureau; Link: http://2010.census.gov/2010census/about/ (last accessed April 14, 2012)

U.S. Census Bureau (2011a). Official Website of the U.S. Census Bureau, 2011 TIGER/Line Shapefiles; Link: http://www.census.gov/geo/www/tiger/tgrshp2011/tgrshp2011.html (last accessed on April 17, 2012)

U.S. Census Bureau (2011b). Official Website of the U.S. Census Bureau, 2011 TIGER/Line Shapefiles, Technical Documentation, Chapter 3; Link: http://www.census.gov/geo/www/tiger/tgrshp2011/TGRSHP2011_TechDoc_Ch3.pdf (last accessed on April 17, 2012)

U.S. Census Bureau (2012). Official Website of the U.S. Census Bureau, Population Estimates; Link: http://www.census.gov/popest/about/index.html (last accessed on April 17, 2012)

Waller, L., Jacquez, G. (1995). Disease Models implicit in statistical tests of disease clustering, Epidemiology, Volume 6, Number 6, November 1995; Link: http://www.jstor.org/discover/10.2307/3703132?uid=3737528&uid=2129&uid=2&uid=70&uid=4&sid=47699075932757 (last accessed on May 11, 2012)

Weather explained (2012). Official Website, Summary of 2002 weather, Review of U.S. events for 2002; Link: http://www.weatherexplained.com/Vol-1/Summary-of-2002-Weather.html (last accessed on June 2, 2012)

Wimberly, M., Lindquist, E., Wey, C. (2011). Analysis of the 2002 equine West Nile Virus outbreak in South Dakota using GIS and spatial statistics, GIS Applications in Agriculture, Volume 3: Invasive species, Boca Raton, FL; Link: http://site.ebrary.com/lib/louisianastate/docDetail.action?docID=10480740 (last accessed on April 8, 2012)

Wordnik (2012). Online Wiktionary, Definition of the word “antigenically”; Link: http://www.wordnik.com/words/antigenically (last accessed on March 21, 2012)