A Spatiotemporal Analysis of Environmental Risk Factors of Lyme Disease in the Northeastern United States

by

Marisa Lynn McGinnis

A Thesis Presented to the
Faculty of the USC Graduate School
University of Southern California
In Partial Fulfillment of the
Requirements for the Degree
Master of Science
(Geographic Information Science and Technology)

December 2018

To my grandmother, Stella McGinnis

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my advisor, Dr. An-Min Wu, for her guidance and encouragement in completing this project. I would also like to thank my committee members, Dr. Jennifer Bernstein and Dr. Elisabeth Sedano for their input during the completion of this thesis, and Dr. Meredith Franklin for her advice. I am grateful to all the professors and staff at the Spatial Sciences Institute of the University of Southern California for all the knowledge and encouragement I have received as I pursue my Master's degree. My family and friends have been very supportive and encouraging throughout this process.

List of Abbreviations

AICc	Akaike Information Criterion
B. burgdorferi	*Borrelia burgdorferi*
CDC	Centers for Disease Control and Prevention
I. dammini	*Ixodes dammini*
I. pacificus	*Ixodes pacificus*
I. ricinus	*Ixodes ricinus*
I. scapularis	*Ixodes scapularis*
ESTDA	Exploratory Spatiotemporal Data Analysis
FIPS	Federal Information Processing Standard
GIS	Geographic Information Systems
GISci	Geographic Information Science
GWR	Geographic weighted regression
MAE	Mean absolute error
MBE	Mean bias error
MRLC	Multi-Resolution Land Characteristics Consortium
NLCD	National Land Cover Database
NOAA	National Oceanic and Atmospheric Administration
OLS	Ordinary Least Squares
RMSE	Root mean square error
SSI	Spatial Sciences Institute
USC	University of Southern California
VIF	Variance Inflation Factor

Abstract

Lyme disease is the most common vector borne disease in the United States. The incidence rate of Lyme disease has been on the rise since it was defined in 1977. From 2000 to 2016, there were over 18,000 cases of Lyme disease diagnosed each year. Of all the confirmed cases of Lyme disease in the United States, 95% occur in the Northeastern and Midwestern states. Lyme disease is contracted by a bite from an infected tick, *Ixodes scapularis*. This research aimed to find the hot spots of Lyme disease and the environmental risk factors, determine the counties that are hot spots in the Lyme disease rate and climate variables maps, and to create a model to test the influence of the variables. Past studies of Lyme disease created risk maps that centered on regression analysis. This study goes a step further to include trend analysis of Lyme disease and the environmental factors while considering spatial and temporal factors.

This study investigated the spatiotemporal trend of the Lyme disease spread rate and environmental factors using hot spot analyses and local Moran's I. A space time cube of these factors was generated and emerging hotspots over 16 years of time period (2000 – 2015) were analyzed. The hot spots were used to identify the correlations of Lyme disease and climate factors. An ordinary least square regression was used to evaluate the relationships between Lyme disease and the environmental risk factors to create an inferential model of Lyme disease. Spatial and temporal environmental risk factors included were precipitation, minimum, mean, and maximum temperature, latitude, longitude, percent forest cover, and year. The variables found to be most significant were year, longitude, latitude, and mean temperature, and explained 14.4% variance of Lyme disease rate in the study area. The significant spatiotemporal environmental factors identified provide researchers and public health officials with updated key factors, and can be used to educate the general public on high-risk areas in the northeastern United States.

Chapter 1 Introduction

In the United States, there have been over 480,000 confirmed cases of Lyme disease between 2000 and 2016. This approximates to 28,000 cases each year from 2000- 2016. As the top 6th Nationally Notifiable Disease ranked by Central for Disease Control and Prevention (CDC) in 2015, Lyme disease is currently the most reported vector borne disease in the United States (CDC 2017b). The incidence of Lyme disease is increasing. While there have been some links between disease hosts (e.g. *Ixodes scapularis)* and environmental factors, the cause of this incidence increase is still unknown. In order to tackle this public health issue, this project aims to understand the spatiotemporal relationships between the Lyme disease incidence and the environmental risk factors.

1.1. Background

To identify the causes of the increased case numbers cases of Lyme disease, it is essential to understand the history and background of Lyme disease. The definition of Lyme disease has changed over the years, and the reporting practices have also spread and changed over time. While these might have directly affected the number of the cases being reported (Kitron and Kazmierczak 1997), the rise in Lyme disease cases is likely more than these reporting inconsistencies. By understanding Lyme disease's history and host cycles, causes of the disease case increase can be established.

1.1.1. History of Lyme Disease

The first medical description of Lyme disease in the United States was not recorded until 1977 by Steere *et al*. (1977). Because it was initially described in a study conducted in Old Lyme, Connecticut, the town was used for naming the disease. Steere et al. (1977) termed it

Lyme arthritis, as it was characterized by swelling and pain in joints. Prior to this, doctors had reported the primary symptom of Lyme disease as a case of erythema migrans, now defined as the initial skin rash at the site of the tick bite (Waller et al. 2007).

While the official definition of Lyme disease has changed over the years, the list of symptoms resembles the original definition. The essential set of symptoms described by Steere et al. (1977) is the same as the classification from the CDC today (2017a, 2017c, n.d.a, n.d.b). These symptoms include erythema migrans, a skin rash at the site of the tick bite, recurrent cases of joint swelling and arthritis, lymphocytic meningitis, Bell's palsy, headache, fatigue, heart palpitations, paresthesia, and a stiff neck. The Lyme disease symptoms in some cases are not serious and will fade in time. On the other hand, they can also result in hospitalization, especially with severe joint arthritis, and might return after a period of remission.

In 1991, the CDC made Lyme disease a nationally notifiable disease. This means that cases are recorded and reported to health departments at the state and local levels for verification, and to document the location of outbreaks (CDC 2017b). As previously mentioned, the case definition provided by the CDC has changed over the years, which in turn allowed for reclassifying symptoms and for the addition of new symptoms. In this study, the CDC definitions from 1996 and 2008 were used in identifying the cases of Lyme disease throughout the study (CDC 2017a, n.d.b). The cases collected from 2000 to 2007 used the 1996 definition; the cases collected from 2008 to 2015 used the updated 2008 definition.

1.1.2. Vectors of Lyme Disease

In 1977, Steere et al. (1977) defined Lyme disease and articulated a set of symptoms. The exact cause of the disease at the time was unknown. It was thought the cause was an arthropod

vector rather than contaminated water. This initial evaluation paved the way for future studies and facilitated the discovery of the vector.

In a progressive study in 1978, it was determined that ticks were correlated to the incidence of Lyme disease (Steere, Broderick, and Malawista 1978). By surveying residents in Connecticut, they found that multiple patients remembered having a tick bite at the site of the lesion, and one patient brought in the *Ixodes scapularis* (*I. scapularis)* tick for identification. Lyme disease is contracted most commonly from a bite from an *I. scapularis* tick infected with the spirochete *Borrelia burgdorferi* (*B. burgdorferi*) (Brownstein, Holford, and Fish 2003). This tick is prevalent in the eastern portion of the United States. Some studies, like Ciesielski et al. (1989), and Spielman (1994) note that there are other species in different geographic regions: *I. pacificus* is a more common vector in the western United States, and *I. ricinus* is a common vector in Europe. Except the known tick species being viable carriers of Lyme disease, there are no other known vectors of Lyme disease (Ciesielski et al. 1989).

The most common tick studied is *I. scapularis*. For many years, *I. dammini* and *I. scapularis* were thought to be two separate species, due to slight differences in feeding and life cycles. However, additional investigations found the two species were slight evolutions of the same species and are now referred to as *I. scapularis* (Oliver 1996; Barbour and Fish 1993). In some cases, *I. dammini*, and *I. scapularis* were used interchangeably. From this point forward, they shall be referred to as *I. scapularis* in this project.

1.1.3. Tick Life Cycle

The life cycle of *I. scapularis* is completed in three stages: larval, nymph and adult stages. Figure 1 below shows the tick life cycle adapted from (Zundel, n.d.).

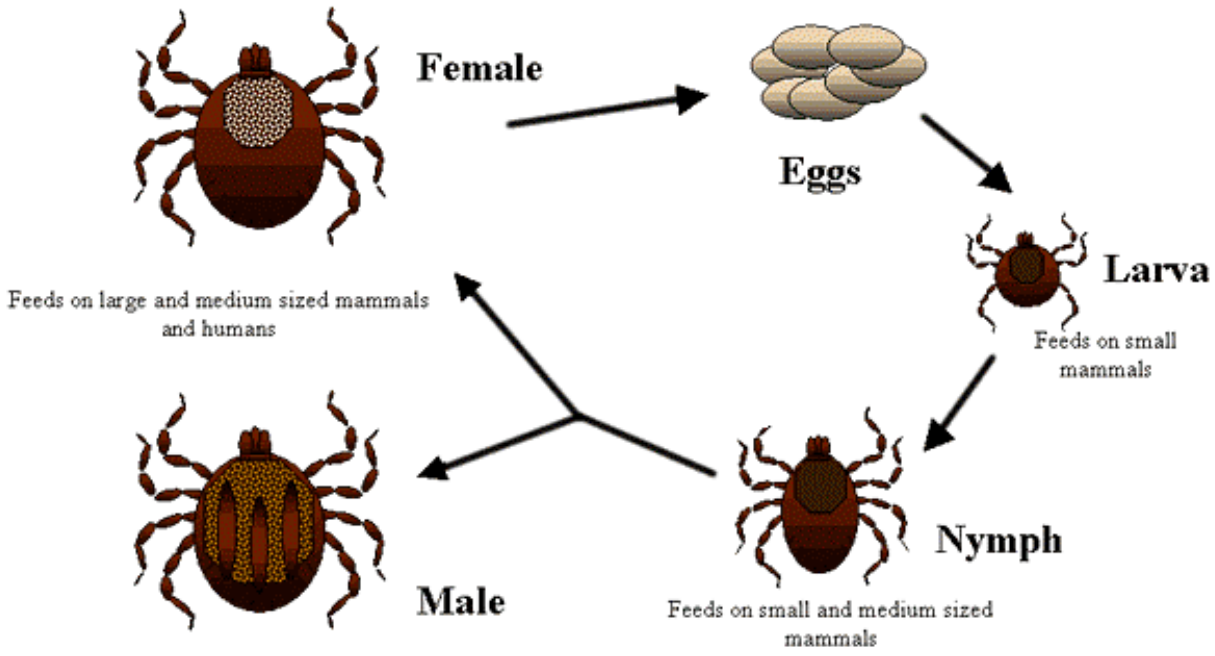

Figure 1. Life cycle of the *I. Scapularis* tick (Adapted from Zundel, n.d.)

Each stage in the tick life cycle requires a continuous blood meal, which lasts anywhere from 1 – 10 days. After the blood meal, the tick falls off from the host and molts. After molting, the tick starts to search for its next blood meal. In a tick's life cycle, about 10% of the time is spent on the host and the rest is spent molting and looking for a new host (Brownstein, Holford, and Fish 2003). Ticks in different stages tend to feast on different hosts. When ticks are in the larval stage, they feed off small animals (e.g. mice). In the nymph stage, ticks feast on small and medium sized animals. In the adult stage, ticks feast on medium and large sized mammals, including humans and household pets.

A tick becomes infected when it feeds on a host infected by a prior bite from an infected tick (Ogden 2010; Spielman 1994). The tick infection is not passed transovarially (transmission to offspring by infection of its eggs); it has to be contracted by feeding on an infected host (Guerra et al. 2002). Thus, Lyme disease can only spread to areas where infected ticks or infected hosts inhabit.

1.1.4. Current Lyme Disease Prevention Methods

The prevention methods for Lyme disease mostly revolve around personal protection. Orloski et al. (2001) notes that utilizing tick repellants, wearing light colored clothes to enhance visibility of ticks, pulling socks over the bottom of pants to protect ticks from crawling up the pant leg, and conducting tick checks when coming in from the outdoors are all ways to help prevent Lyme disease. While these methods would help with noticing a tick, and hopefully catching and removing the tick before being bitten, prevention methods could definitely be improved.

Without accurately knowing where Lyme disease occurs, and what causes the increase of Lyme disease cases, it is difficult to create effective prevention methods (Brownstein, Holford, and Fish 2003). As suggested by Barbour and Fish (1993) and Orloski et al. (2001), the second common method of Lyme disease prevention, besides personal protection, is vector control. In order to create effective prevention methods to completely help control the vectors, it is critical to have a concrete understanding of how the ticks are affected by the environment, climate, and other variables.

1.2. Study Scale

The majority of cases of Lyme disease in the United States are reported in the Northeastern and Midwestern states. It is important to note, however, that all 48 contiguous states and Alaska have reported cases of Lyme disease in the past. In 2015, 95% of the confirmed cases were reported from Maine, Vermont, New Hampshire, New York, Massachusetts, Connecticut, Rhode Island, Pennsylvania, New Jersey, Maryland, Delaware, Virginia, Wisconsin and Minnesota (Waller et al. 2007; CDC 2017a). Because this is a sizable area, the spatial extent of the study area for this project was scoped down to five Northeastern

states. The study area consists of Maine, Vermont, New Hampshire, New York, and Massachusetts for their high case counts and contiguity. The study area is shown below, in Figure 2.

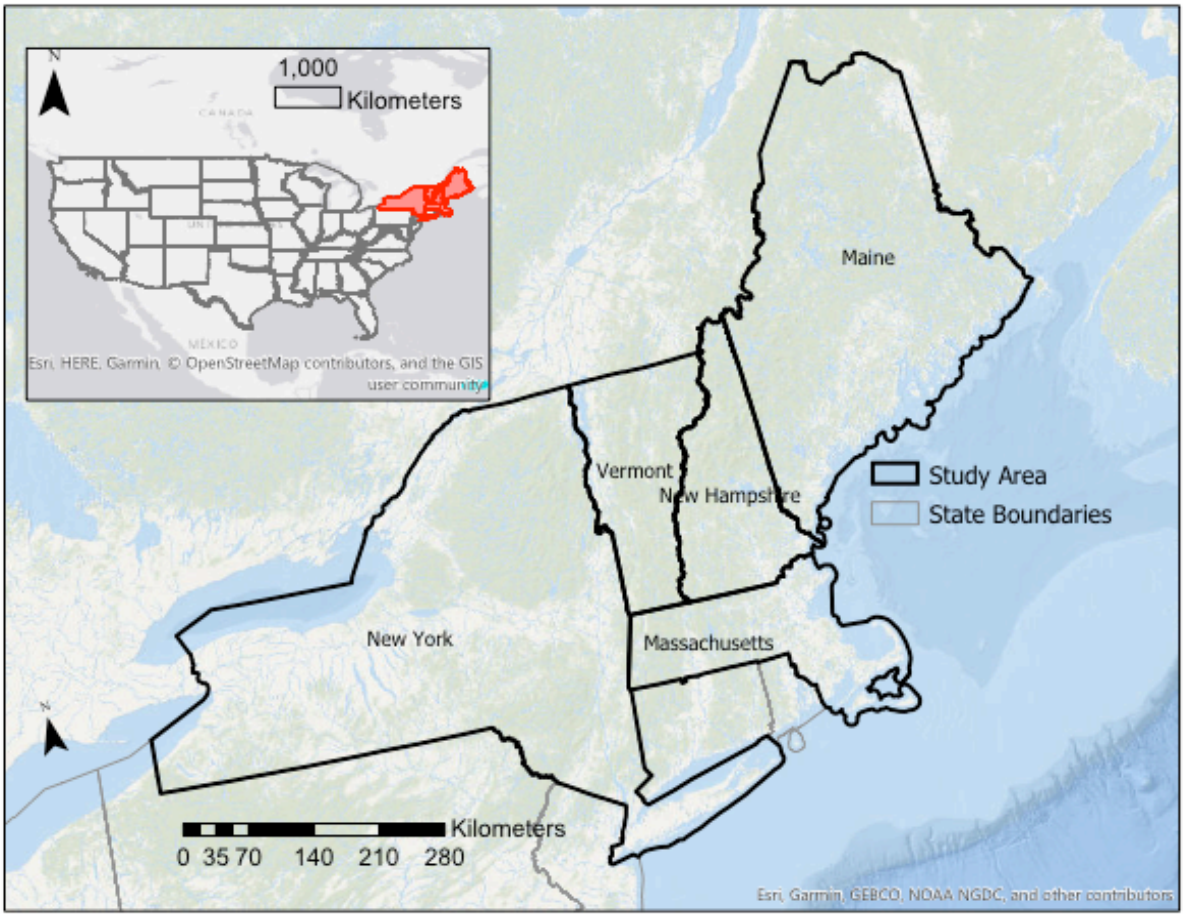

Figure 2. The Study Area includes five states in the Northeastern United States

Another important aspect of the study is its temporal scale. The CDC provides yearly counts of Lyme disease cases at the county level from 2000 to 2016. However, the climate data used for the study was only available for the years of 2000 – 2015 when the data was collected. The time span of this study thus matches the data availability for both Lyme disease cases and the environmental factors for 2000-2015.

The number of Lyme disease cases varies spatially and temporally. For the study area, there were a total of 153,486 cases of Lyme disease over 16 years (2000-2015), with 4,000 being reported to the state level only. Within the total of 116 counties, the average is 80 cases per county per year, and 1285 cases total per county over the 16 years. The maximum count was 1720 cases for Dutchess County, New York in 2002, and the minimum count value was zero in multiple counties (CDC 2017a). This spatial and temporal variability in the number of cases makes further analysis of the relationship between the environmental variables and cases of Lyme disease possible.

1.3. Motivation

The motivation of this study is to combine Geographic Information Science (GISci) and epidemiology to increase the understanding of the causes and risk factors of Lyme disease. Spatial analysis and spatial statistics allow for identification of habitat factors and can then be mapped against populations to predict potential risk for disease outbreaks (Kitron 1998; O'Sullivan and Unwin 2010). GISci also offers tools to analyze and identify the most important factors in the spatial and temporal distributions of *I. scapularis* and the spirochete *B. burgdorferi*.

The changes in *I. scapularis* population lead to the Lyme disease rate change over time. The ability to analyze the spatial and temporal distribution is integral in clarifying the relationships between the Lyme disease vector *I. scapularis* and the risk factors. Once there is a clear understanding of why and how the *I. scapularis* varies during the year and over the years, prediction and prevention efforts can utilize this information (Brownstein, Holford, and Fish 2003). For *I. scapularis*, 90% of the life cycle is spent off the hosts, which means the climate and environment would have an increased significance and effect on the tick. Because of this, climate

was chosen as one of the factors to be analyzed in this study. To further represent the available habitat areas for ticks, the percent of forest cover is also included in the analysis of this study.

1.4. Research Objectives

The aim of this study was to investigate the spatiotemporal relationships between Lyme disease rate and environmental factors. To accomplish this goal, there were three main research objectives:

1. Determine the locations of the hotspots of Lyme disease cases per 100,000 and of the environmental factors both spatially and temporally;
2. Determine environmental hot spots that correlate to the Lyme disease hot spots in the county level; and
3. Create a model to test the influence of the selected environmental factors on the Lyme disease incidence rate.

To achieve the first objective, a space time cube was created and an emerging hot spot analysis is conducted. The hot spots of the data, as well as the trend over the years, were identified. The effect of using a rate instead of a count for Lyme disease was demonstrated by creating maps for both the Lyme disease rate and count. The local Moran's I was also included to note the spatial clusters and outliers in each of the datasets.

The second objective was executed by visually surveying the hot spot maps and observing the counties associated with multiple hot spots. This provided valuable insight on the relationships between the hot spots of Lyme disease and the environmental factors. The result of the visual correlation elucidated the relationship of an environmental factor with Lyme disease rate. In some cases, this correlation may not be what was anticipated.

The third objective was accomplished by an ordinary least squares (OLS) regression. It was conducted to analyze the correlations between Lyme disease rate and environmental factors including precipitation, temperature, and forest coverage. In addition, a geographic weighted regression (GWR) was tested as it accounts for local influences in the data. The result from the OLS regression identified the environment factors that were significant and explanatory to the variance of the Lyme disease rate. The OLS regression coefficients were used to create an inferential model of Lyme disease rate. The data for all 16 years is used to prepare the model and identify the significant variables, and a residual analysis was included to determine the model performance.

Data provided by the CDC (2017a; 2017b) contained case counts for each county in the United States using the definition of Lyme disease at the time of year. The cases of Lyme disease included in the dataset are cases diagnosed during the collection period. Previously, Waller et al. (2007) used the Lyme disease data from the CDC from 1990 to 2000 for a risk map. This research provided an update of the Lyme disease data use and included a trend analysis as well. The environmental factors are expanded upon from the Waller et al. study for the investigation of the Lyme disease rate. This project also further advances the study of Lyme disease by creating an inferential model to further examine the relationships of Lyme disease.

In terms of environmental factors, the Lyme disease rate was compared to the annual minimum, mean, and maximum temperatures of the year, the annual precipitation and the percent of forest cover per county. Research by Brownstein, Holford, and Fish (2003, 2005), and Ogden et al. (2010), has linked temperature and land use to tick abundance and cases of Lyme disease. The forest cover variable is included to represent land use, specifically tick inhabited areas.

1.5. Layout

The structure of this thesis includes five chapters, each developed based on the previous chapter. Chapter 2 details previous studies on Lyme disease, the current risks of contracting Lyme disease, the known populations with the highest risk, and the prevention methods related to Lyme disease contraction. Chapter 3 contains the data descriptions and the methodology used to analyze the data to achieve the results. Chapter 4 describes the results of the analysis and includes the relevant discussions related to the results. The final chapter, Chapter 5, contains the conclusions of the study as well as its limitations and future research investigations.

Chapter 2 Related Work

Chapter 2 seeks to present relevant information on Lyme disease to support this study. The combination of epidemiology and ecology with GISci allows for the spatial patterns of Lyme disease incidence to be identifiable. The strength and accuracy of models and analysis of Lyme disease are stronger in cases where Geographic Information Systems (GIS) and epidemiology are combined (Glass et al. 1995). The full range of risk factors and causes of Lyme disease are still largely unknown, and further study is needed to understand them. The use of GISci allows for analysis and further understanding of the spatiotemporal trends. Goodwin et al. (2001) noted that the temporal fluctuations have a great deal of influence on the availability of ticks. The climate and other environmental factors also fluctuate over time, so it is important to determine the variables most significant factors for tick survival. When they are determined, an accurate prediction model can be created.

2.1. Transmission of Lyme Disease

The transmission of Lyme disease is dependent on the vectors of Lyme disease. In Northeastern United States, the vector is the *I. scapularis* tick. The life cycles of the tick, along with the exposure to *B. burgdorferi* spirochete (a spiral bacterium), influence the cases of Lyme disease. The transmission of *B. burgdorferi* generally occurs from the animal hosts that the young ticks feed on, and rarely through tick birth (Guerra et al. 2002). This means that ticks start off uninfected, become infected while feeding on infected animals, and then spread infection to their future hosts. A tick only needs to feed from an infected host once to become an infectious agent. Conversely, the disease can carry over from one stage of the tick life cycle to the next (Gilmore, Mbow, and Stevenson 2001).

The occurrence of Lyme disease also depends on the number of ticks available. The infection rate of ticks is influenced by the number of larval ticks from the previous years, the available hosts, and the tick survival in between blood meals (Goodwin, Ostfeld, and Schauber 2001). Thus, the number of cases is higher in years where tick survival is unhampered. The climate factors are thus significant to understand tick survival in between blood meals.

In essence, human infection depends on the tick prevalence and the exposure to ticks in outdoor areas. While there is the opportunity to contract Lyme disease when in an endemic area, Ogden et al. (2010) note that when bitten by ticks that were not engorged, the victim was less likely to contract Lyme disease. This is believed to be caused by the lack of a blood meal on an infected host prior to an individual being bitten.

2.1.1. Current Risk

The analysis of the risk of contracting Lyme disease is problematic, with multiple complications affecting surveillance. The most widespread problem is the changing definition of Lyme disease. The definition of Lyme disease has changed over time to include additions from new research, and to limit the number of misdiagnoses or overdiagnoses (Waller et al. 2007). Additionally, CDC (2017c) noted that the increase of Lyme disease cases in an area might be the product of something other than a true increased incidence of cases. Common increases of cases result from the spread of reporting practices, updated reporting practices, reporting bias, and location of diagnosis (Waller et al. 2007). This results in uncertainty about the causes of the increased incidence.

The other common problem with identifying the risk of contracting Lyme disease is the difference in locations between where the patient is diagnosed and where the patient is infected. Human travel is a typical cause of this uncertainty in geographic distribution (Brownstein,

Holford, and Fish 2003). Individuals may travel as short distances such as across their town, or further distances such as crossing multiple counties in a day. The CDC dataset used in this study is based on locations at time of diagnosis (CDC 2017b). The analysis based on this data will be contingent on probability of contracting Lyme disease based on diagnosis location.

2.1.2. At Risk Population

While individuals can contract Lyme disease at any age, people in certain demographics are more likely to contract Lyme disease. Orloski et al. (2000) studied a dataset that had samples with the age ranged from under 1 to 100 years old; the mean age was an average of 35-39 years old. Steere et al. (1977) and the CDC (2017b) noted more cases occurring in children. When the samples were broken down by gender, the prevalence was slightly over 50% for males. Orloski et al. (2000) noted 51% of the patients were men, Ciesielski et al. (1989) found 53% were male, and Davis et al. (1984) found 54% were male. The difference in contraction rates of the males and females in these studies was not substantial. Thus, males were not significantly more likely to contract Lyme disease.

The main noticeable correlation between the noted demographics is the likelihood of spending time outdoors. The important condition in contracting Lyme disease is being in a habitat that supports *I. scapularis*. Goodwin, Ostfeld, and Schauber (2001) and Barbour and Fish (1993) note that human infection is contingent upon humans residing in an area where host-seeking ticks live. Working outdoors, whether as an occupation or recreation, is more common among men, while being outside recreationally would support the correlation between young children and adults in their 30s. To help estimate the areas where Lyme disease can be contracted, the forest cover percent per county was included in the regression variables. Counties

with a higher percentage of tree canopies are hypothesized have a greater area where ticks and their hosts could be located.

2.1.3. Tick Hosts

The tick hosts are an important part of the tick life cycle, and the spread of Lyme disease. The most common host a tick first feeds from is the white-footed mouse (Guerra et al. 2002; Goodwin, Ostfeld, and Schauber 2001; Waller et al. 2007). Another common host is the white tailed deer. Ogden et al. (2010) brought up an excellent point, that while wild animals are often tick hosts, cats, dogs and even other humans are also common tick hosts. While domesticated animals may carry Lyme disease, it is easier to study wild animals when testing ticks for Lyme disease due to the increased exposure to ticks.

Birds are another a key host population. They have an important role in transporting opportunistic ticks from one area to another (Spielman 1994). While deer are able to transport ticks as well, birds have the unique ability to carry the ticks to a non-native place and start the cycle of Lyme disease there. The spread of Lyme disease has occurred in this manner, shown through the cases now occurring in Quebec, Canada and even farther north, with more spread and prevalence believed possible in the future years (Ogden et al. 2010; Brownstein, Holford, and Fish 2005).

It is noteworthy that, while many animals may act as hosts for the ticks, not all of them transfer Lyme disease. Animals that capable of carrying the Lyme disease pathogen and able to transmit the pathogen to a tick are termed competent reservoirs (Barbour and Fish 1993). White-tailed deer, white-footed mice, birds, and even lizards are all common hosts for ticks, but they are all not competent reservoirs. The southern *I. scapularis* feed on lizards, which function as a host for ticks, are not competent reservoirs. This results in lower rates of Lyme disease in the

southern states. The incidence of Lyme disease is higher in the northeastern United States because of the prevalence of suitable hosts (Barbour and Fish 1993).

2.2. Environmental Factors

The use of environmental risk factors is routine when studying Lyme disease cases and causes. The major risk factors evaluated in the different studies include temperature values and temperature aggregates, precipitation, land use, and soil type. Glass et al. (1995) linked humidity, temperature, slope, and land use to tick abundance. The 1998 study on vector-borne diseases by Kitron (1998) lists forest cover, sandy soil, and hardwoods as significant factors, based off of the needs of the tick hosts.

Guerra et al. (2002) and Brownstein, Holford, and Fish (2003; 2005) use minimum and maximum temperatures and precipitation as main factors for Lyme disease specifically. Brownstein, Holford, and Fish (2003) supported the use of temperature, and note specifically that minimum temperature was the only variable to have a simple positive relationship with cases, which is believed to show the lower limit of tick habitat survivability. The maximum temperature was correlated to Lyme disease cases (Brownstein, Holford, and Fish (2003). Ogden et al. (2010) utilized temperature alone, while Waller et al. (2007) used land use, soil type, and moisture as their main variables. Ogden et al. (2010) included data reported at the weather stations, and interpolated the data to have an average temperature dataset, while Guerra et al. (2002) used precipitation and average temperature data available from the National Oceanographic and Atmospheric Administration (NOAA).

Besides climate variables, land use was also investigated in the course of this study. The consideration of land use as an environmental risk factor was multifaceted. The National Land Cover Database (NLCD) contains data on land use and change for the contiguous United States,

and includes useful analysis tools (Homer et al. 2007). Investigation into the rate of land use change by the Multi-Resolution Land Characteristics Consortium, also MRLC, (2017) show that there is less than 3% change in the United States overall from 2001-2011. This supports what Homer et al. (2007) noted in their report on the NLCD. In an effort to utilize land use data, without using the categorical data available in the NLCD, Kitron (1998), and Nicholson and Mather (2014) use the percent of forest coverage as an estimate of land use. Nicholson and Mather used forest cover to estimate the areas where nymph ticks would consistently inhabit, and to determine suitable habitats.

2.3. GIS and Lyme Disease

Geospatial techniques can be used to investigate spatial and temporal trends of public health datasets such as the Lyme disease. GIS allows data from different sources and formats to be merged spatially, and analyzed as a whole (Szwarcwald et al. 2000). In the case of Lyme disease, the use of GIS allows various environmental factors to be included and analyzed against population and cases of Lyme disease (Glass et al. 1995; Kitron 1998). This can be used to show risk, and create predictions based on the data available.

2.3.1. Past Studies

Various studies undertaken on Lyme disease covered a range of data types and analysis methods. Many of the spatiotemporal studies used point data to study the cases or the vectors of Lyme disease (Ogden et al. 2010; Goodwin, Ostfeld, and Schauber 2001; Guerra et al. 2002; Steere et al. 1977; Glass et al. 1995). On the other hand, several studies used aggregated data, for both vectors and the cases (Orloski et al. 2000; Kitron and Kazmierczak 1997; Waller et al. 2007; Ciesielski 1989). The studies that used aggregated data created risk maps on case counts and conducted habitat suitability analyses.

The spatial distribution of the pathogen *B. burgdorferi* is integral to increased incidence of Lyme disease. Unfortunately, the spatial variation of the pathogen is only available by studying ticks, *I. scapularis,* to determine the number of infected ticks and the number of infected tick hosts. Glass et al. (1995) note that when analyzing a large study area (at the state or region level), the use of environmental factors is acceptable to use in analysis rather than the study of ticks and the host populations. This is due to the high cost of time and money spent when analyzing the hosts. Thus, studying the cases per county per year is adequate when combined with the study of environmental factors.

Waller et al. (2007) note there was evidence of a relationship between the increase in incidences of Lyme disease and the expansion of reporting practices and the refinement of the CDC definition of Lyme disease. Moreover, when analyzing Lyme disease cases, a major concern is the number of misdiagnoses, lack of diagnoses, reporting bias, and imprecise serological results, which means the number of confirmed cases in the CDC data may be less than the actual incidence rate (Kitron and Kazmierczak 1997; Waller et al. 2007). This indicates the presence of false negatives, and the process of verification through the CDC allows only confirmed or probable cases to be counted in the data counts per year.

Most of the Lyme disease studies had focused their area of study in the northeastern United States, where the ticks are prevalent and the majority of cases are reported. Guerra et al. (2002), LoGiudice et al. (2005) and Waller et al. (2007) studied multiple states in the Northeastern United States and Midwest. Studies by Glass et al. (1995) and Frank et al. (2002) focused on one county or state in the Northeastern United States. In Ogden et al. (2010), the authors looked at the border of the Northeastern United States and Canada. Other studies looked at the United States as a whole. Orloski et al. (2001) and Ciesielski et al. (1989) studied the USA

as a whole, utilizing data available from the CDC. The study scales were dependent on the elements studied, whether it was the cases of Lyme disease or the vectors, and the availability of data. In instances where the data was self-collected, the study areas tended to be much smaller.

It is important to note that each Lyme disease study is unique for the area and the scale it studies. While conducting Lyme disease studies at a large spatial scale is important in order to learn more about the disease in a specific area, there is the likelihood of an inherent fallacy being assumed in the study results. In the study by Brownstein, Holford, and Fish (2003), the authors note that the probability surface of Lyme disease for the contiguous United States had a lower suitability in the West Coast. They inferred the difference in the Pacific region vector was the cause of lower suitability; this vector was declared as the *I. pacificus* tick.

2.3.2. Risk Maps

A common analysis result for Lyme disease study is a risk map. In two studies, the risk maps were created by classifying the selected risk factors, performing logistic regressions, and evaluating the results by the chi-squared test (Waller et al. 2007; Glass et al. 1995). The concern with risk mapping is there is no additional trend analysis included in the study. The risk maps only demonstrate the probability or risk of contracting Lyme disease based on the factors. In two separate studies utilizing aggregated data, the results of the study consisted of a risk map, with no further analysis included (Kitron and Kazmierczak 1997; Waller et al. 2007). The risk map created in the 1997 study by Kitron and Kazmierczak (1997) is shown below in Figure 3. The map shows the endemicity of Lyme disease based on the occurrence of ticks in the wildlife surveys, and Lyme disease cases. The risk maps in Waller et al. (2007) were based on the annual crude incidence rates for the counties.

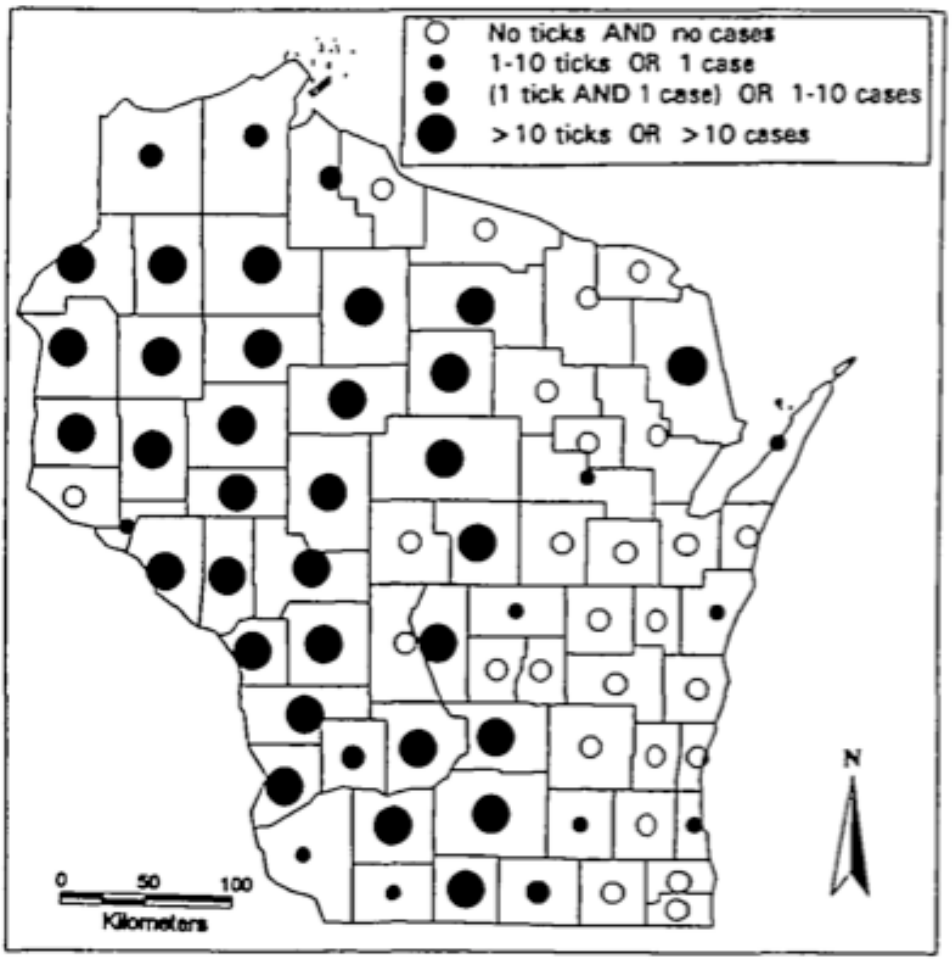

Figure 3. Sample risk map from study by Kitron and Kazmierczak (1997)

2.3.3. Data Types and Challenges

The Lyme disease data types used in Lyme disease studies and GIS are either point, polygon, or non-spatial (table form). For studies using data points of individual Lyme disease cases, the data was generally collected in the course of the study (Brownstein, Holford, and Fish 2003; Guerra et al. 2002; Goodwin, Ostfeld and Schauber 2001). There are some exceptions, like Ciesielski et al. (1989) and Glass et al. (1995) who used point data published by the CDC. The point datasets by the CDC have been discontinued, and the current dataset available is a non-spatial table with cases per county (CDC 2017b). This dataset was used in Waller et al. (2007). Other studies, such as Kitron and Kazmierczak (1997), Davis et al. (1984), and Orloski et al.

(2000), used Lyme disease data aggregated to the county boundaries (polygon form) from local health departments.

When looking at Lyme disease data aggregated to county polygons, the scope of the study is more regional. Kitron and Kazmierczak (1997) focused on using spatial statistics and GIS to correlate the county polygon aggregation Lyme disease data with the tick distribution, the Lyme disease case distribution, the percent of wooded areas, and the human population density. Waller et al. (2007) look in the Northeastern United States at the county aggregation level as well to compare climate and ecologic variables to county incidence, and created a risk map based on the analysis results.

The major challenge in the study of Lyme disease is the lack of individual data. The reported data is often aggregated. Geographic aggregation is an established method of making data unidentifiable, while still allowing the data to be available for analysis. By utilizing aggregated data, the cases of Lyme disease cannot be tied back to any one individual and therefore the location of any individual. With the United States information health privacy laws, like the Health Insurance Portability and Accountability Act (HIPAA), the use of health data includes the need for de-identified data. Aggregation of the data is critical when using sensitive information, like health data and census data (Amrhein and Reynolds 1997). The benefit of using an aggregate dataset with limited accuracy of locations is the inherent privacy for sensitive data provided by aggregation (Longley 2012).

Jelinski and Wu (1996) noted that there are ways to utilize aggregation, and one particular method was to highlight the rates of change in spatial analysis. This is accomplished by looking at the incidence rate instead of the case count, and evaluating the trend over time. The main concern of this is ecological fallacy, an error to solely infer the result of the analysis done

in the aggregated data to an individual (Longley 2012). When using a time series of data at the county level, the rate of change in Lyme disease cases can be evaluated by observing data standardized by the population of that county. Kitron and Kazmierczak (1997) and Waller et al. (2007) standardized the data by looking at cases of Lyme disease per 100,000 persons per year at the county level. The standardized rate allows for a more accurate analysis of the true incidence and correlation. Another aspect of using aggregated data is sometimes it is the only available data (Amrhein and Reynolds 1997). If there are no other options, the use of aggregated data is better than no data at all.

When analyzing different datasets, the data must be in equivalent units. Holt, Lo, and Hodler (2004) note that it is important for data to be analyzed at the same areal unit, using normalization or areal interpolation to be compatible. The equivalence of data is needed to ensure that analyses are being properly done; otherwise the results of analysis are misleading. The use of areal interpolation to estimate data points is quite common with weather data, as the weather is not measured systematically. Instead, the data collected at the weather stations are averaged and interpolated to create a raster surface (Ogden et al. 2010; Brownstein, Holford, and Fish 2005; Guerra et al. 2002). In these studies, the use of areal interpolation for the weather data worked well, allowing for fairly accurate study outcomes. The climate surfaces can be used in analyses that require a single value per polygon by using summary statistics to determine a value for each polygon or unit (Brownstein, Holford, and Fish 2003). Reed et al. (1993) note that when using larger quadrats there are stronger correlations with the ecological aspect being measured.

The results of the investigation of past Lyme disease studies in this section influenced the creation of the methods of this thesis. Many studies have been conducted on Lyme disease at varying aggregations, risk factors, and scales. Using the aggregated Lyme disease case data (e.g.

cases per county), data standardization is needed to accurately compare data across areal boundaries. A common method of standardizing Lyme disease cases is by the population, shown as cases per 100,000 persons (Waller et al. 2007). This allows for the case prevalence to be compared equally across areas with varying populations.

The limitations of the past studies were taken into account to create this study. Important aspects included the study area, the environmental factors selected, how to best use the aggregated data and the study scale. These studies created the precedent that this study was based on. This study used aggregated Lyme disease data in conjunction with climate and land use to create a predictive model for the Northeastern United States at the county level.

Chapter 3 Methods

The goal of this study is to understand the distributions of Lyme disease cases and their environmental risk factors over space and time in the Northeastern United States. As stated in Chapter 2, previous investigations have created risk maps in this region. The intent of this study was to take the risk maps a step further, and include a spatiotemporal trend analysis to elucidate the spatiotemporal trends of this public health threat, and to determine the influence of the environmental factors in an influential model for the Lyme disease risks. The full cause of the increase of Lyme disease cases is still unknown. Correlation between Lyme disease and environmental factors has shown there is a relationship between the two.

By analyzing the environmental risk factors for Lyme disease, the relationship between Lyme disease cases and the environment was elucidated. This chapter describes the data and the methods used for achieving this goal. The methods described were completed using Microsoft Excel for table formatting and Esri ArcGIS Pro 2.1.0 for the spatial analyses. Hot spot analysis and local Moran's I analysis were conducted on the cases of Lyme disease and the environmental factors to help understand the data trends and the correlation between Lyme disease and the factors. The Lyme disease model was then created using a stepwise least squares regression.

3.1. Research Outline

The research design flowchart in Figure 4 demonstrates the overall analysis flow in this study. There are two separate analysis processes. The first portion of the investigation (noted as the 'Setup' process in Figure 4) includes data acquisition and compilation, an extensive data exploration using spatial statistics, and building the space-time cube. The second portion is the spatiotemporal analysis that consists of emerging hot spot analysis, local outlier maps using local Moran's I, the overall trend of Lyme disease, and a comparison of the identified hot spots. The

final portion was the creation and verification of an inferential model of Lyme disease rate using stepwise regression and residual analysis. Both OLS and GWR were tested to find the best-fitted model. The residuals from the model were mapped for the study region and a supplementary residual analysis was conducted to analyze the performance of the final model. A full explanation of the methods is included in the rest of the chapter.

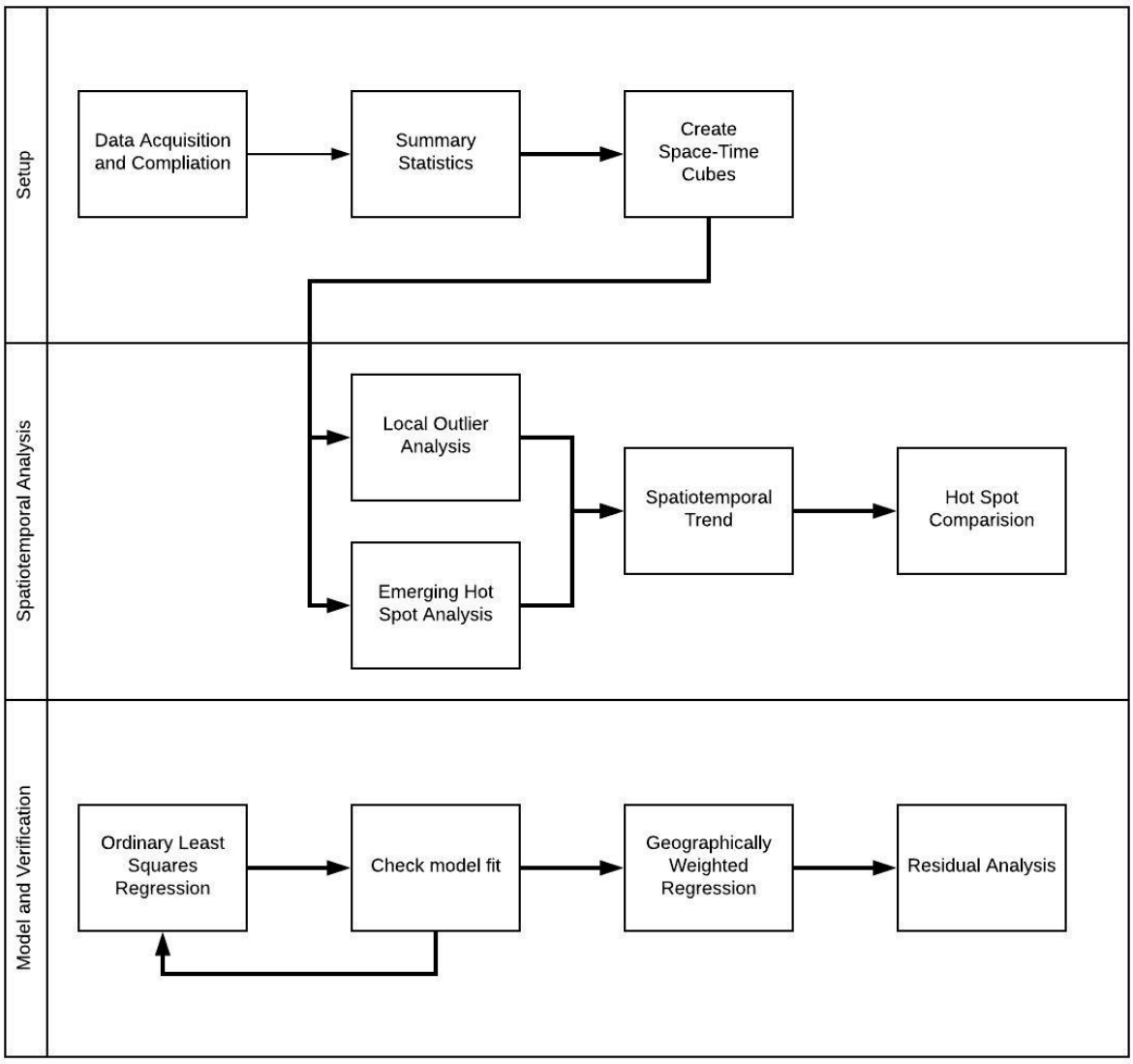

Figure 4. Research design flowchart

3.2. Data Acquisition and Compilation

There are four datasets required to conduct this study; the Lyme disease case counts, the yearly climate data, the forest cover percentage and the county subdivisions. The datasets fell into three categories: shapefile, table, and raster. The county boundaries were a polygon shapefile that included attributes of some demographic data. The Lyme disease cases and the climate data were in table forms. The forest cover data was a raster dataset, which required extra processing to be properly formatted before analysis.

3.2.1. Lyme Disease

The Lyme disease data is assembled at the county level and published by the CDC annually. The dataset downloaded from the CDC website for this study was a .csv file containing the cases of confirmed Lyme disease in each county of the study area from 2000 to 2016. The dataset contained County Name, State Name, State Code, County Code, and the yearly case counts. The 2016 data column was ultimately deleted because of no corresponding climate data in that year. The state and county codes were joined together to create the Federal Information Processing Standard (FIPS) for each county. Table 1, below, shows a portion of the Lyme disease dataset. The data is precise to the reporting year but not to the date of contraction. Another caveat for this dataset is that the cases were reported from where they were diagnosed, not necessarily where they were contracted. While there is apparent uncertainty in this dataset, this is the best available Lyme disease data found during the research period for the region.

Table 1. Excerpt of the Lyme Diseases Cases Table

Ctyname	Stname	STCODE	CTYCODE	Cases2000	Cases2001	Cases2002
Androscoggin County	Maine	23	1	0	1	2
Aroostook County	Maine	23	3	0	1	1
Cumberland County	Maine	23	5	13	14	46
Franklin County	Maine	23	7	0	0	2
Hancock County	Maine	23	9	0	6	5
Kennebec County	Maine	23	11	2	0	7
Knox County	Maine	23	13	4	1	7
Lincoln County	Maine	23	15	3	2	14
Oxford County	Maine	23	17	5	2	1
Penobscot County	Maine	23	19	2	6	10
Piscataquis County	Maine	23	21	0	1	1
Sagadahoc County	Maine	23	23	0	2	3
Somerset County	Maine	23	25	0	1	2
Waldo County	Maine	23	27	0	2	1
Washington County	Maine	23	29	1	1	2
York County	Maine	23	31	40	68	112

Source: Centers for Disease Prevention and Control 2017a

The Lyme disease cases dataset required limited processing. The original coverage of this dataset was the entire United States of America. It was selected for the necessary records, and then the extraneous records were deleted so it only contained records for the study area. Next, a new column was inserted to contain the County FIPS. The equation used to calculate this was

$$FIPS = STCODE \times 1{,}000 + CTYCODE \quad \text{eq. 1}$$

This created the proper 5-digit County FIPS code as the unique identifier for the county record.

3.2.2. County Divisions

In a geospatial study, the most important dataset is the spatial component, which includes the area and boundaries of the study area. The county boundaries data was a polygon shapefile downloaded from the US Census Bureau. It contains demographic data such as the population for 2010. The Census Bureau products have a 95% confidence interval (United States Census

Bureau 2016), so the data is as accurate as can be directly obtained and is ready to use as is after download.

The county boundaries data has the spatial extent for the entire United States. The study area of the five northeastern states was selected and exported to a new shapefile (Figure 3). Except the required attribute fields - GeoID, County Name, and State Name, and the population total for 2010, all extraneous columns were deleted from the attribute table. The GeoID field represents the FIPS in a string type and allows the other datasets to be easily joined to this spatial dataset. This is identical to the FIPS in the Lyme disease data.

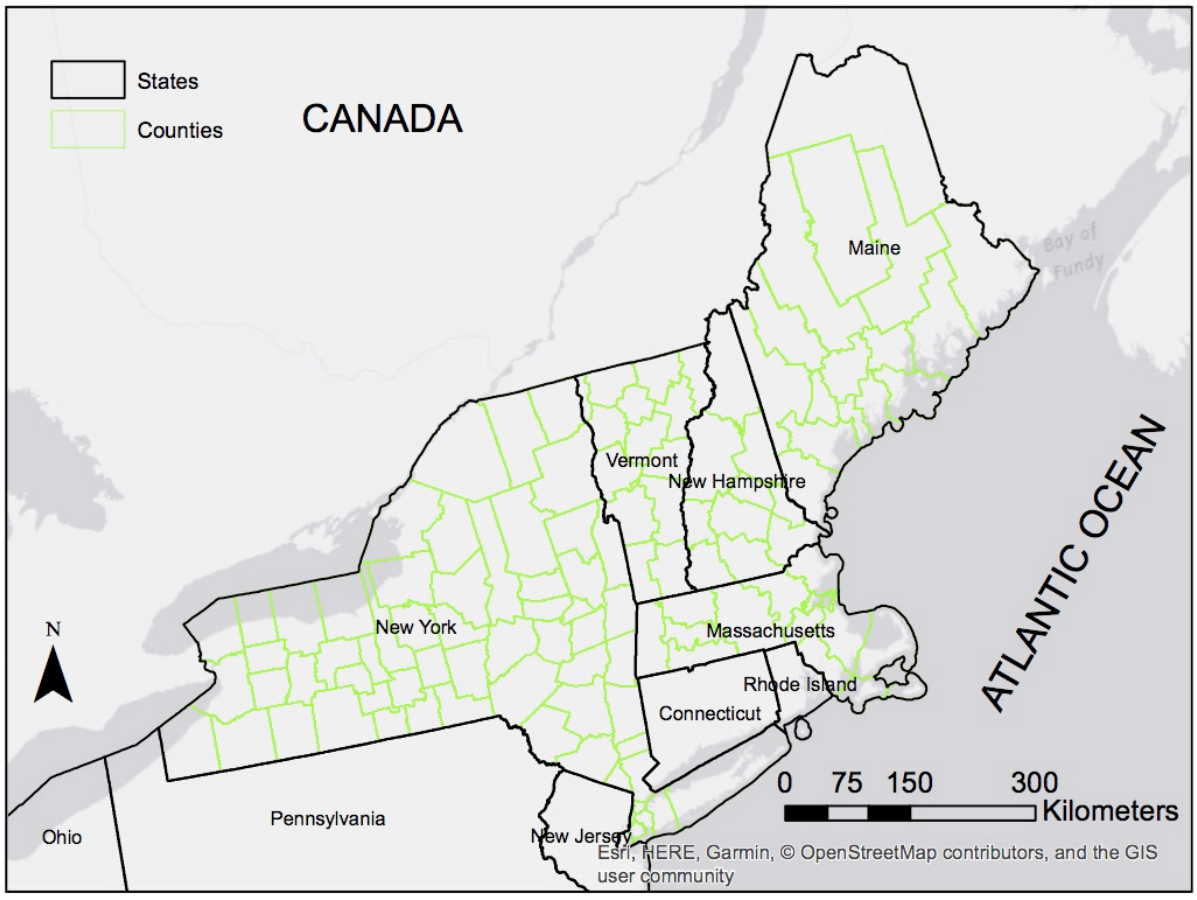

Figure 5. County boundaries of the study area

In the attribute table of the county boundary data, the centroids of the longitude and latitude were calculated. After this was calculated, the shapefile was projected into USA Contiguous Albers Equal Area Conic. This projected coordinate system was selected for use throughout the entire research as it preserves area and avoids areal distortion during projection. As the study area spans multiple latitudes, this coordinate system also appears to be appropriate.

3.2.3. Climate

Having climate data for the county level for 16 years was integral in this project. The Parameter-elevation Regressions on Independent Slopes Model (PRISM) Climate Group was selected as the data source because it publishes climate data at many levels, including the county level. As a subdivision of Oregon State and the Northwest Alliance for Computational Science and Engineering, PRISM Climate Group gathers historical climate observations from various sources and developed short- and long-term spatial climate datasets for the United States (PRISM Climate Group, 2018). The data required for this study was acquired through the data explorer located on the PRISM website (http://prism.oregonstate.edu). At the time of download, the most current data available was 2015, so the temporal extent of the study was restructured to the period between 2000 and 2015.

The data explorer on the PRISM website allows data download by coordinates with various criteria settings. The precipitation data was recorded in inches, and the temperature data was recorded in degrees Fahrenheit. According to PRISM, the county centroid coordinates were appropriate estimates for use (PRISM Climate Group, 2018). In the data explorer, the Annual Values criteria were set from 2000 to 2015 and the climate variables including precipitation, minimum temperature, maximum temperature, and mean temperature were selected. The PRISM

data was then acquired by a .csv file exported from the county attribute table in ArcGIS Pro, which included the counties' centroid coordinates (latitude and longitude) and the FIPS code.

The acquired climate dataset included the counties' centroid coordinates (latitude and longitude) and the FIPS code and required a few changes before being able to use in Lyme disease modeling. First, the first 10 rows of the data table were metadata and were removed so the dataset included only the data. Moreover, the climate dataset had one row per county per year. In order to comply with the requirement for building a space time cube (see Section 3.3.2), a Date column was created with the date string in the format of 12/31/YEAR (YEAR being the last four digits of the year). Table 2 shows an example section of the final climate table:

Table 2. Excerpt of the Edited Climate Table

Name	Longitude	Latitude	Elevation	Year	Date	ppt	tmin	tmax	tmean
23019	-68.6494	45.4005	318	2000	12/31/2000	41.21	31.1	52.2	41.6
23019	-68.6494	45.4005	318	2001	12/31/2001	26.34	32	54.9	43.5
23019	-68.6494	45.4005	318	2002	12/31/2002	39.74	32	52.9	42.5
23019	-68.6494	45.4005	318	2003	12/31/2003	48.1	30.7	52.1	41.4
23019	-68.6494	45.4005	318	2004	12/31/2004	39.14	30.9	52.2	41.6
23019	-68.6494	45.4005	318	2005	12/31/2005	67.66	32.3	53.5	42.9
23019	-68.6494	45.4005	318	2006	12/31/2006	55.56	35.1	55.1	45.1
23019	-68.6494	45.4005	318	2007	12/31/2007	51.14	30.6	52.7	41.7
23019	-68.6494	45.4005	318	2008	12/31/2008	55.37	31.9	53.1	42.5
23019	-68.6494	45.4005	318	2009	12/31/2009	50.73	31.3	52.4	41.8
23019	-68.6494	45.4005	318	2010	12/31/2010	51.91	36.1	55.8	46
23019	-68.6494	45.4005	318	2011	12/31/2011	52.93	33.4	54.1	43.7
23019	-68.6494	45.4005	318	2012	12/31/2012	47.8	34.5	55.2	44.9
23019	-68.6494	45.4005	318	2013	12/31/2013	48.89	33.1	53.4	43.2
23019	-68.6494	45.4005	318	2014	12/31/2014	54.9	32	52.6	42.3
23019	-68.6494	45.4005	318	2015	12/31/2015	47.06	31.7	52.8	42.2

Source: PRISM Climate Group

When looking at the data range of the data, there was an abnormality. For Barnstable County, Massachusetts, and Wayne County, New York, the climate variables were negative for

all years. The temperature values were set to -17966.2 °F, and the precipitation was -393.66 inches. Upon further investigation, these values were found to be the null values input from PRISM.

To estimate the missing values in the PRISM dataset, two models were created and the predictions were compared with the PRISM data for accuracy. When evaluating the models on a test site, the variance between the models and the Prism data was less than .5 degree Fahrenheit, so the model outputs were deemed acceptable. The Prism data used for the estimation was a 30m raster for each year, and each climate variable.

To start, the two counties with missing values were selected and exported into individual shapefiles. The models were exactly the same, with the addition of the first model projecting the climate rasters to the USA Contiguous Albers Equal Area Conic coordinate system. The second model used the projected raster created in the first model. Multiple models were used to prevent the creation of multiple copies of the projected rasters. Figure 6 shows the model without the raster projection. In the models, the county shapefile was used to clip the raster to the county boundaries, and then the Calculate Statistics tool was used to find the mean of the raster cells in the county boundary. This was repeated for each climate variable and each year. The model outputs were manually transferred to the Climate dataset to replace the null values. This allowed for the proper values to be used during the remainder of the analyses.

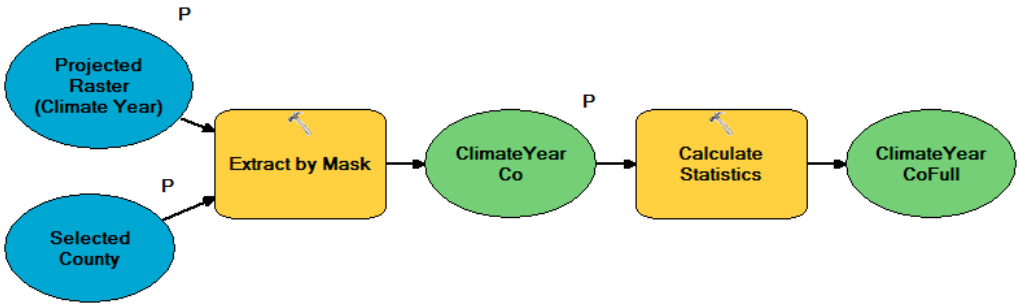

Figure 6. Climate Estimation Model for Missing Data

3.2.4. Forest Cover Data

With the hosts of Lyme disease in mind, forest cover was used to represent the potential land cover area available for the Lyme disease contraction. The NLCD contains various land cover products established by the MRLC. The Forest Cover dataset was chosen from the NLCD for its numerical data type. This quantitative data characteristic lent itself to working with regressions. The NLCD 2001 Percent Tree Canopy was the version selected for all data regardless of the year. The data methods were updated in between updates, so the 2001 and 2011 datasets were not compatible. Thus, there was spatial variation in the dataset, but no temporal variation. The forest cover data was deemed adequate for use due to the spatial variation between the counties in the study area.

Downloaded as raster datasets for individual states, the NLCD datasets were merged as a mosaic dataset before projecting into the USA Contiguous Albers Equal Area Conic projection. The raster data represented the percent forest cover in a 30 m spatial resolution. This is the same spatial resolution as the PRISM rasters. To estimate the percent of forest cover for each county, the Zonal Statistics as a Table tool was used to provide the county polygons as zones with the zone field set as the FIPS code. This estimation was created and run using Modelbuilder in ArcGIS Pro. The output of the Modelbuilder was a table with all the standard statistics

calculated, and the mean, the averaged percent of forest cover for each county was the variable used for further analysis (Section 3.4.1).

Another estimation model was tested before finalizing the model above. In this model, the counties included a 30-kilometer buffer (approximately 18.6 miles), to approximate areas of forest cover that the population may have traveled. The buffered forest cover area was divided by the only the area of the county, because if it was divided by the area of the whole buffer the percent forest cover decreased. This resulted in some counties with a percent forest cover over 100%, so was ultimately decided against.

3.2.5. Data Processing

Before analysis and exploration began, the data was properly formatted and joined. The Transpose Fields tool was used on the Lyme disease data to combine the 16 separate year columns into one column named Year. The result was 16 rows per county, each with a column of the Lyme disease count and a year column representing the year of Lyme disease being reported.

The second phase of data processing and preparation was to join the newly formatted Lyme disease data and the climate data. This was accomplished by creating a unique identifier for each row, a combination of the FIPS and the year. A new field was added in both tables, and the values were calculated by adding the two fields in one string. The FIPS-Year column was calculated as:

$$UniqueID = FIPS \times 10{,}000 + Year \qquad \text{eq. 2}$$

The Add Join tool was then employed to join the two tables based on the unique yearly ID. The result was a table with 16 rows per county, each with the cases of Lyme disease, the year of the disease reported, and precipitation, minimum temperature, maximum temperature, and mean temperature for each year.

The next phase of data processing and preparation was to join the data table to the County shapefile. This was accomplished using the Spatial Join tool. In ArcGIS Pro, the tool allows a one to many join. The FIPS was used as the common field to join the Lyme disease and climate data to the County Shapefile. Later in the study, the join field tool was used to join the mean percent forest cover field to the county shapefile, with the FIPS as the identifier.

After the data was joined, the Lyme disease rate (the disease count per 100,000 county population) was calculated in a new column. The equation used to calculate this was:

Lyme Disease Rate (%) = [Disease Counts / Population] x 100,000 eq. 3

This created a normalized value for Lyme disease. By normalizing the values, it allowed for a more equivalent comparison across varying county populations.

3.3. Spatiotemporal Data Analysis

The spatiotemporal data analysis section consists of exploratory data analysis, and trend analysis. The exploratory data analysis was to ascertain if the data chosen was adequate for this study and to provide a deeper insight on the data variation. The trend analysis provided an in-depth look at the Lyme disease rate and the environmental factors before the modeling. The set of data exploration methods included summary statistics, the creation of space time cubes followed by an emerging hot spot analysis, and a local outlier analysis for both the Lyme disease rate and the climate variables.

3.3.1. Summary Statistics

The summary statistics were used to determine if the data was normally distributed before the regression analysis. This set of statistics was included to document the data properties for Lyme disease and the included variables. This was added after the data analysis had started and therefore no transformation was performed for any non-normally distributed data.

The summary statistics investigated the histogram, the skewness and kurtosis, and the values of mean, median, and standard deviation. The histogram indicated the frequency distribution of the data values, with an expected normal distribution as a bell shaped curve. The mean and median indicate the center of distribution of the data, with the average value and the median value, respectively. The data distribution is more normal when the mean and median values are close together. The kurtosis indicated how likely the data distribution would produce outliers. In a normal distribution, the kurtosis is close to or equal to 3. The skewness showed the symmetry of the data, with an ideal value of 0.

3.3.2. Space Time Cube

Space time cube was used for visualizing spatial data over a span of time in ArcGIS and ArcGIS Pro. Building space time cubes was integral to this study as it was utilized in a variety of tools, including the Emerging Hot Spot Analysis tool, the Local Outlier Analysis tool, and the Visualize the Space Time Cube in 2D tool. One space time cube was created with each of the variables, including the Lyme disease count, rate, and the climate variables, to visualize the variable's spatial and temporal trend.

The space time cube layout is quite unique, as it contains space and time information at the same time. Figure 7 shows an example of the file format. The X and Y axes show the spatial location, while the Z axis, labeled Then to Now, represents the time periods. Each spatial location for each year has its own distinct bin to hold the values. Bins that are for the same spatial location will have the same location ID. The file format is a netCDF file.

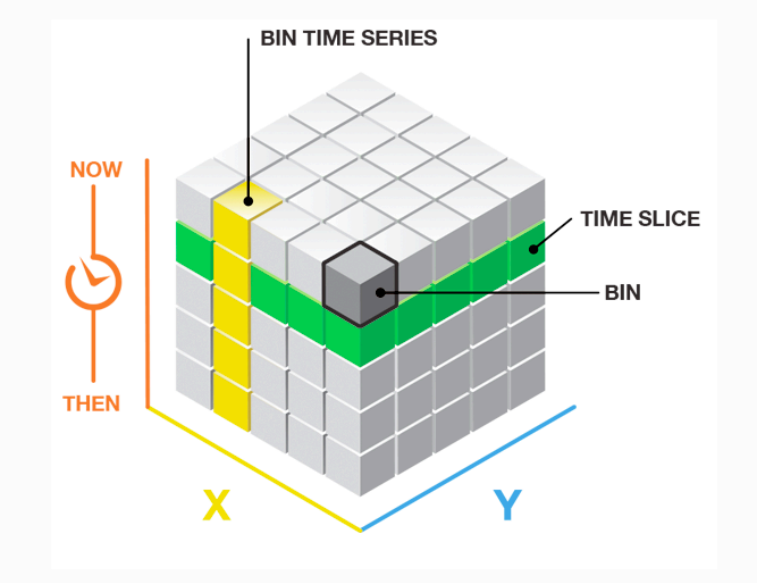

Figure 7. Example of a space time cube (adapted from Esri, 2018d).

Every variable of this study with spatial and temporal series of data available was included in one space time cube created using the Create Space Time Cube From Defined Locations tool in ArcGIS Pro. Variables with no temporal variations, like longitude, latitude, and percent forest cover, could not be included in the space time cube analysis. The time step interval was 1 year, as the interval for the Lyme disease data and the climate data is one year. As the data was already temporally aggregated, there was no temporal aggregation in the space time cube. One of the mandatory settings was available fill, which was set to fill empty bins with space time neighbors for the areas missing values. This is a mandatory field to be set before the tool can run; however the data investigation and background statistics ensures there are no missing values.

3.3.3. Emerging Hot Spot Analysis

The space time cube created from Section 3.3.2 was first applied to evaluate temporal and spatial trends using the Emerging Hotspot Analysis tool in ArcGIS Pro. This tool calculated the Getis-Ord Gi* statistic for every variable per county per year, and evaluated the Gi* statistic test results (hot spot and cold spot trends) using the Mann-Kendall trend test. The z-scores and p-

values from the Getis-Ord Gi* statistics determine the counties where high or low values spatially accumulate. The Mann-Kendall trend test assesses the temporal values to determine the overall trend for each county (Esri 2018a).

The result of the Emerging Hot Spot Analysis displays the spatial and temporal trend in each county in a quantifiable and understandable approach. This is imperative as it shows the areas that are statistically different from their surroundings. In this case, it highlights the areas with higher or lower Lyme disease rates, Lyme disease counts, precipitation, temperature, and percentage of forest cover. The emerging hot spot analysis classified the cells into eight different types of hot or cold spot categories, except for the areas with no statistically significant patterns detected. The categories of the emerging hot spot and their definitions can be seen in Table 3.

Table 3. Emerging hot spot classifications used in this study

Classification	Definitions
No Pattern Detected	Does not fall into any of the hot or cold spot patterns defined below.
Consecutive Hot Spot	A location with a single uninterrupted run of statistically significant hot spot bins in the final time-step intervals. The location has never been a statistically significant hot spot prior to the final hot spot run and less than ninety percent of all bins are statistically significant hot spots.
Intensifying Hot Spot	A location that has been a statistically significant hot spot for ninety percent of the time-step intervals, including the final time step. In addition, the intensity of clustering of high counts in each time step is increasing overall and that increase is statistically significant.
Persistent Hot Spot	A location that has been a statistically significant hot spot for ninety percent of the time-step intervals with no discernible trend indicating an increase or decrease in the intensity of clustering over time.
Diminishing Hot Spot	A location that has been a statistically significant hot spot for ninety percent of the time-step intervals, including the final time step. In addition, the intensity of clustering in each time step is decreasing overall and that decrease is statistically significant.
Sporadic Hot Spot	A location that is an on-again then off-again hot spot. Less than ninety percent of the time-step intervals have been statistically significant hot spots and none of the time-step intervals have been statistically significant cold spots.
Oscillating Hot Spot	A statistically significant hot spot for the final time-step interval that has a history of also being a statistically significant cold spot during a prior time step. Less than ninety percent of the time-step intervals have been statistically significant hot spots.
New Cold Spot	A location that is a statistically significant cold spot for the final time step and has never been a statistically significant cold spot before.
Persistent Cold Spot	A location that has been a statistically significant cold spot for ninety percent of the time-step intervals with no discernible trend, indicating an increase or decrease in the intensity of clustering of counts over time.
Sporadic Cold Spot	A location that is an on-again then off-again cold spot. Less than ninety percent of the time-step intervals have been statistically significant cold spots and none of the time-step intervals have been statistically significant hot spots.

(Source: ESRI 2018a)

The emerging hot spot analysis was conducted on Lyme disease data as well as the environmental factors. The emerging hot spot tool was run on both the cases and cases per 100,000 population to show the importance of data normalization. The climate variable included

in the space time cube (precipitation and maximum, mean, and minimum temperature) were also tested for the emerging hot spots to view the temporal trend. Contiguity edges and corners was chosen for the conceptualization of spatial relationships setting in the emerging hot spot analysis it ensured at least one neighbor would be included in the analysis.

3.3.4. Local Outlier Analysis

The second use of the space time cube was to study the local spatial autocorrelation via the Local Outlier Analysis tool. The Local Outlier Analysis tool calculated the local Moran's I statistic to see if there are any spatial clusters or dispersion (outliers) over a time span. Similar to the Getis Ord Gi* statistics, the local Moran's I identifies statistically significant clusters of high or low values. In addition, it also identifies spatial outliers in the data (Esri 2018b). This analysis was used as additional data explorations because it could identify not only hot spots, but also unexpectedly high or low values that contradicted the values in the surrounding counties. As the analysis involves both spatial and temporal distributions, the input required a space time cube generated as a netCDF file in Section 3.3.2. The Local Outlier Analysis identified 6 different trends shown in Table 4.

Table 4. Local Outlier Analysis Classifications

Type	Definition
High - High Cluster	Statistically significant cluster of high values
Low - Low Cluster	Statistically significant cluster of low values
High - Low Outlier	Statistically significant high value surrounded by low values
Low - High Outlier	Statistically significant low value surrounded by high values
Multiple Types	Multiple types in one county
No Statistical Significance	No statistical trend

Source: Esri 2018b

The definitions of the resulting six categories from Anselin local Moran's I are also included in Table 4. The first category is the High-High Cluster, meaning a county of high values is surrounded by the counties of high values. The second category is Low-Low Cluster, meaning a county of low value is surrounded by the counties of low values. The third category is High-Low Outlier. The High-Low Outlier category means a county of high value is surrounded predominantly by counties with low values. The opposite is the Low-High Outlier category, where a county with a low value was surrounded by counties with mostly high values. When there are multiple types of statistically significant cluster and outliers occur in the same county over time, it is categorized as Multiple Types. The final category is No Statistical Significance, where no statistical significance of clusters or outliers is detected.

As with the emerging hot spot analysis, this local outlier analysis was run on Lyme disease cases, rate, and the environmental factors. The specifications for the tool were the same for all versions. The number of permutations was set at 999, as the smallest possible pseudo p-value is 0.001, the standard in statistical studies. The conceptualization of neighbors was set

again to contiguity edges and corners. This negated the need for a distance band but ensures at least a neighbor included in the analysis.

3.3.5. Space-Time Visualization and Optimized Hot Spot Analysis

Using the Visualize Space Time Cube in 2D tool, the overall trend of the Lyme disease incidence was evaluated. This tool was established based on the results of space time cube used for both of the emerging hot spot analysis and the local Moran's I. In this study, it was used to demonstrate the comprehensive spatiotemporal trend of the Lyme disease rate and the Lyme disease count. The final analysis conducted in this study was an Optimized Hot Spot Analysis of the forest cover variable. This tool was an alternative hot spot analysis needed for the percentage of forest cover, as the forest cover data did not contain a time series. The tool identified hot spots throughout the study area of the same time period using the Getis-Ord Gi* statistics. The Optimized Hot Spot Analysis tool resulted in p-values and the confidence intervals for the hot and cold spots.

3.4. Regression Modeling

The OLS regression and GWR were tested for creating the Lyme disease model. The modeling was included as an additional analysis to determine the significant variables of Lyme disease, and to determine the nature of the relationships. The OLS regression analysis has the ability to handle spatiotemporal datasets using a time-series attribute (e.g. year) and/or coordinates (i.e. latitude and longitude). The GWR tested for local spatial variations in subregions of the study area. For the purpose of identifying the relationships between the Lyme disease rate and the spatial and temporal variables, both OLS and GWR were kept in linear models. The models were both evaluated against a set of model criteria to determine the model fit and quality.

3.4.1. Ordinary Least Squares Regression

The OLS was employed to examine the relationship between Lyme disease and the environmental variables. Lyme disease was the dependent variable and the potential independent variables include precipitation, minimum temperature, maximum temperature, mean temperature, year, percent forest cover, longitude, and latitude. These environmental variables were tested to see whether and how much they could explain the variance of the Lyme disease cases in the Northeastern USA. The year and longitude and latitude were tested to see if there was a spatial or temporal trend. All the factors will now be referred to as independent factors due to the inclusion of time and location variables in the regression. The regression equations are as follows:

$$y = \beta_0 + \beta_1\chi_1 + \beta_2\chi_2 + \ldots + \beta_n\chi_n \qquad \text{eq. 4}$$

where y represents the dependent variable Lyme disease rate, X_n represents independent variable *n*, and β_n is a partial coefficient of each independent variable *n*. The coefficient reflects the relationship between the independent and the dependent variables. The result of this tool was an output feature class that contained dependent variable estimates and residuals, and a report. These outputs were important in determining the accuracy of the model.

The OLS Regression model was processed in a stepwise regression. The model was run and the results evaluated. If the model did not sufficiently satisfy the model checks, another regression was conducted. The following regression would remove the variable with the highest insignificant probability and with a variance inflation factor (VIF) over 7.5.

The regression was initially tested with precipitation, average minimum temperature, mean temperature, and average maximum temperature as independent variables. While the minimum and maximum temperature were selected specifically for tracking tick survivability,

the mean temperature was included in the regressions to observe the climatic trend. Precipitation was included to estimate host survival. With climate as a known strong predictor in Lyme disease cases, the climate trends might have affected the case incidence. The regression was run and then rerun, with the elimination of one variable whose partial coefficients were not statistically significant (i.e. p-value > 0.05) and the p-value was the highest. This process was repeated until all partial coefficients of the independent variables were significant.

It was determined that more variables were needed to enhance the model performance based on the model results. The variables selected were county centroid longitude (X) and latitude (Y), percent forest cover for each county, and year of Lyme disease cases reported. The centroid coordinates were included to add a spatial context, while the year was included to add a temporal context. The forest cover for the counties was included to provide an insight on the availability of areas where there could be exposure to the vectors of Lyme disease in each county. These new variables were chosen to fill the gaps and to increase the accuracy of the model.

The stepwise regression was attempted again with the new variable set. The set included year, longitude, latitude, annual precipitation, annual minimum temperature, annual mean temperature, annual maximum temperature, and forest cover. The stepwise regression was conducted with two procedures (as follows). The first time the variables were eliminated based on the variable or variables with the highest VIF and then the tie decided by the highest probability (p-value). In the second round, the variables were eliminated solely based on which had the highest partial coefficient p-value.

The regression model outputs were evaluated for the models. The first check was the sign (+ or -) of the coefficients, to see whether the influence of the variable on the disease rate was in

the expected direction. The larger coefficients showed a stronger relationship, while the smaller coefficients show a weaker relationship (Esri 2018c). The second check was the variable redundancy (or multicollinearity). This was evaluated by the VIF. Following the guideline in ArcGIS Help (Esri 2018c), the threshold of VIF was set to 7.5. This prevents redundant variables from being included in the model. If the variables explain similar portions of the model, they have a higher VIF. The model with VIF values equal or greater than 7.5 is likely to have the independent variables highly correlated to each other and thus only the models with VIFs less than 7.5 were acceptable. The third check was to see if the variables have a statistically significant coefficient. The p-value was set to 0.01, thus having a confidence interval of 99%.

The next three checks were to determine the overall fit and quality of the model. The fourth portion to verify was the significance of the Jarque-Bera statistic. This represented whether the regression residuals were normally distributed to satisfy the regression assumption. If the residuals were normally distributed, they over-predict and under-predict equally. The next check determined the model performance, represented by the Akaike's Information Criterion (AICc) and adjusted R-squared values. The AICc represented the fit of the model in respect to model complexity. Low AICc were preferred. The adjusted R-Squared represented the amount of variance was explained by the independent variable. Thus, a higher adjusted R-Squared value was preferred for the higher variances of Lyme disease explained by the overall model. The sixth check was to test for spatial autocorrelation in regression residuals. This was to evaluate whether there are spatial patterns remaining in residuals and if model specification is appropriate. Finally, based on the model evaluation results, the best-fit Lyme disease models were chosen.

Among all of the model criteria, the Koenker test also served as an indicator of whether GWR should be applied. The Koenker test showed if the relationship between Lyme disease and

the independent variables was a stronger prediction in some areas and not in others. A significant Koenker test suggests GWR would be better suited instead of OLS regression.

While GWR was planned to generate a prediction model for the Lyme disease incidence, the model was unable to be effectively established. Independent variables could not be selected into a single GWR model because of the high multicollinearity between longitude, latitude, and year. The environmental factors also exhibited multicollinearity. While the model could be run successfully with individual independent variable, GWR was determined not viable for this study. As the result, the predictive model would instead be created based on the OLS regression.

3.5. Residual Analysis

A residual analysis was conducted on the best-fitted regression models for model performance. The residual analysis included residual plots, statistics of the residuals, and mapping the residuals. The statistics of the residuals included the Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and the Mean Bias Error (MBE). The equations used for these statistics were:

$$RMSE = [N^{-1} \Sigma (P_i - O_i)^2]^{0.5} \quad \text{eq. 5}$$

$$MBE = N^{-1} \Sigma (P_i - O_i) \quad \text{eq. 6}$$

$$MAE = N^{-1} \Sigma |P_i - O_i| \quad \text{eq. 7}$$

where N is the total number of samples, with P_i and O_i being the predicted and original dependent value, respectively. As the name indicates, RMSE takes a square root for the sum of the squared residuals (the difference between the predicted value P_i and the original value O_i) then averages this value to determine how large the overall residual of the model is. The RMSE of zero (0) would indicate a perfect model fit without any error. An MAE is the measure of the average absolute difference between the observed values (O_i) and the predicted values (P_i). The

ideal absolute value for this would also be 0, indicating a perfect model fit with the average difference between observed and predicted as 0. The final statistic, MBE, gave the overall bias of the model. It is important to note that the positive and negative difference can cancel out, displaying the model uniformly over- or under-estimation. The ideal value for the MBE is 0.

Chapter 4 Results and Discussion

This chapter presents the results of spatiotemporal analysis and modeling for Lyme disease cases and its relationships with environmental factors in Northeastern states. Discussions regarding the analytical results were also included. The organization of this chapter follows the similar structure in Chapter 3 as the analyses proceeded: Section 4.1 covers the summary statistics for Lyme disease and the explanatory variables. Section 4.2 includes the results and discussion of Lyme disease hotspots, local outliers and emerging hotspots for spatial and temporal trends of Lyme disease rates. Section 4.3 contains the results and discussion of hot spots of the environmental factors and compares them to the locations of Lyme disease hot spots. The comparison of the hot spots lends additional insight on the correlation between Lyme disease and the environmental factors. This chapter closes with the Lyme disease model and subsequent verification in Section 4.4.

4.1. Summary Statistics

The summary statistics of Lyme disease and the independent variables were included to be the foundation of the data investigation. Table 5 contains the overall summary statistics, including the mean, standard deviation, median, kurtosis, and skewness. The last column denotes if the data has a normal distribution.

Table 5. Summary statistics for regression variables

Name	Mean	Standard Deviation	Median	Kurtosis	Skewness	Normal?
Lyme Count	80.33	155.21	15	80.33	3.82	No
Lyme Rate	52.73	106.46	15.74	54.55	5.62	No
Precipitation	47.09	8.87	47.06	3.61	0.36	Yes
Minimum Temp	36.82	4.60	36.40	3.42	0.51	Yes
Mean Temp	46.47	3.38	46.20	3.33	0.24	Yes
Maximum Temp	56.11	3.36	56.10	3.15	-0.10	Yes
Forest Cover	51.25	22.10	56.12	2.40	-0.63	No
Longitude	-73.49	2.65	-73.61	2.42	-0.13	No
Latitude	43.07	1.21	42.99	2.85	0.16	No
Year	2007.50	4.61	2007.50	1.79	0	No

The majority of the independent variables were normally distributed as expected. Of the five potential independent variables (precipitation, minimum temperature, mean temperature, and maximum temperature, and percent forest cover), only percent forest cover variable was not normally distributed. The longitude, latitude, and year variables were also not normally distributed. The examples of the histograms for precipitation and forest cover can be seen in Figure 8 and Figure 9. Figure 8 illustrates the overall normality of precipitation and Figure 9 exemplifies a non-normal distribution of forest cover.

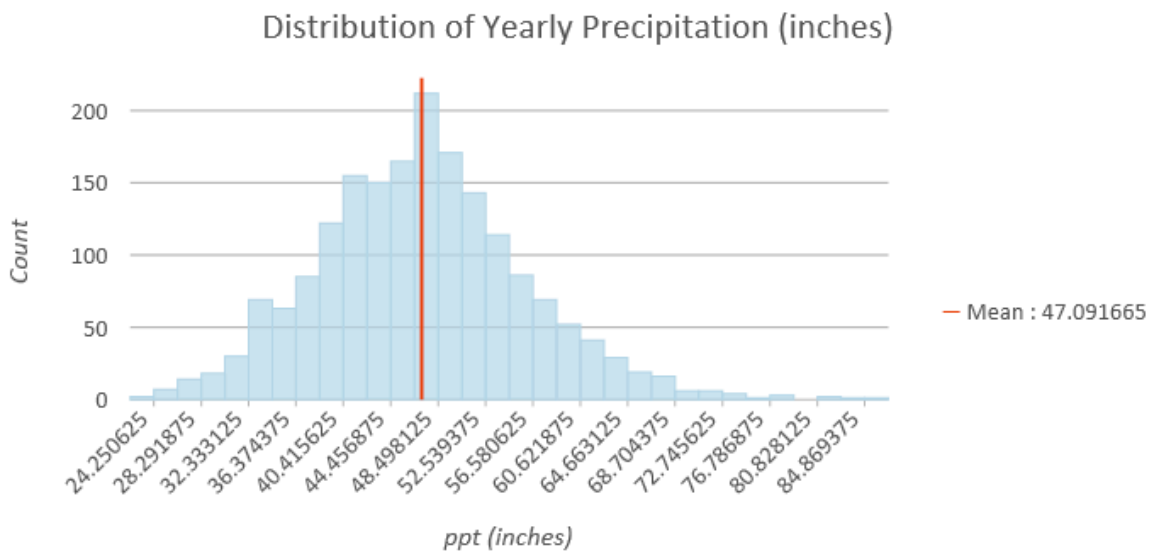

Figure 8. Precipitation histogram showing the normal distribution of the data

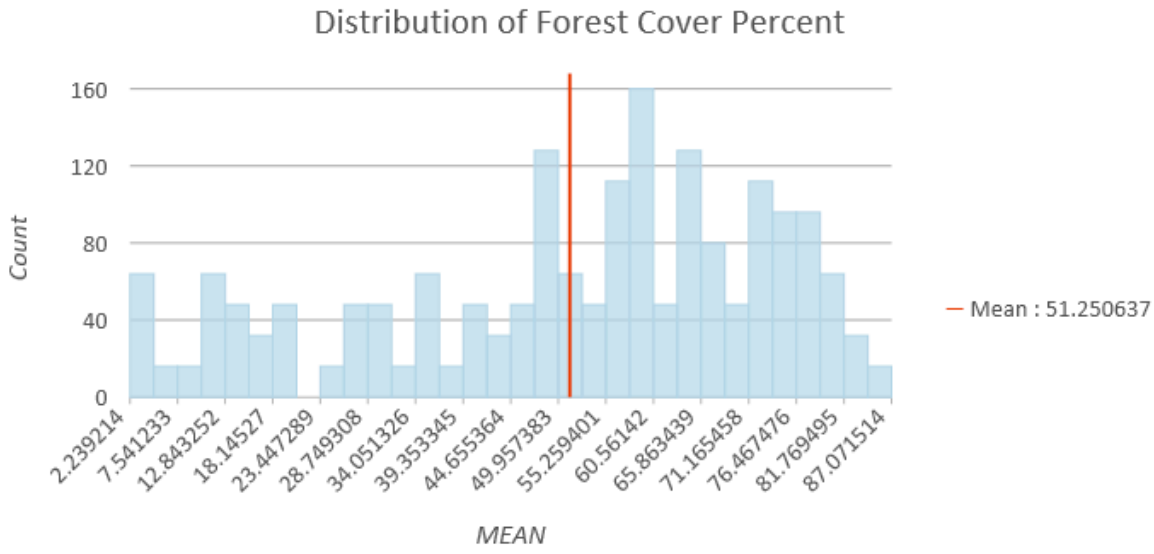

Figure 9. Forest Cover histogram showing the data is not normally distributed

The distribution of the Lyme data was also investigated in the summary statistics table. Both the rate and the count of Lyme disease were not normally distributed. It was not expected that the Lyme disease data would have a normal distribution, due to the variance in case counts and rate throughout the study area. For further insight on the data variance, scatterplots for 2010, 2014, and 2015 were created (Figures 10, 11 and 12, respectively).

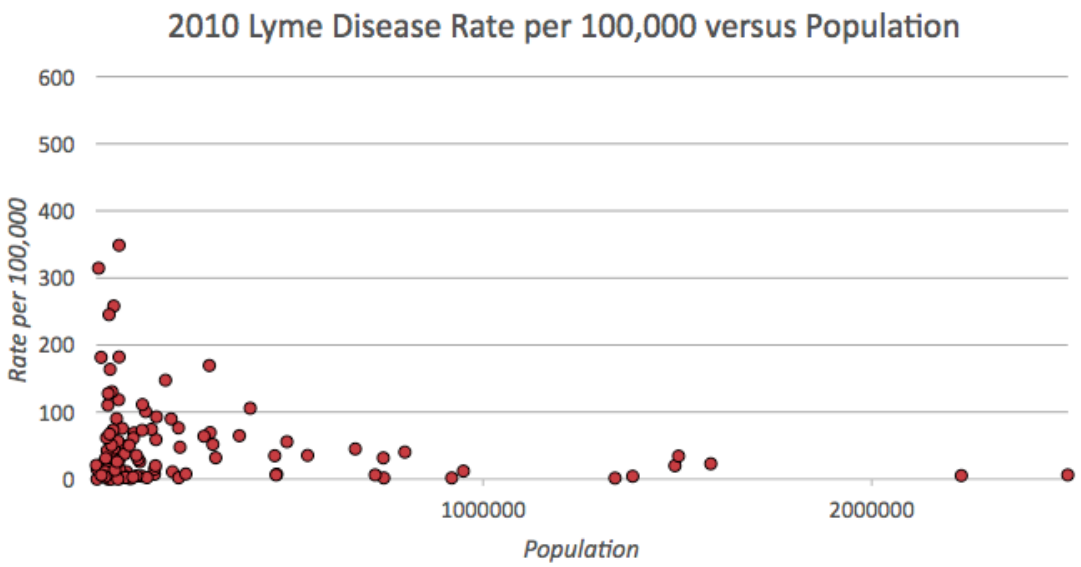

Figure 10. Scatterplot for 2010 Disease rate

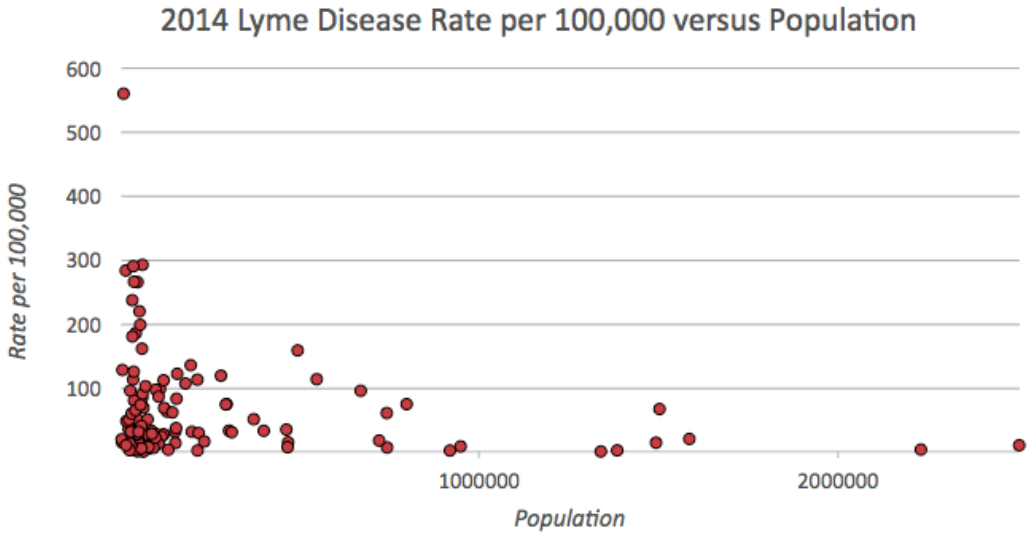

Figure 11. Scatterplot for 2014 Disease Rate

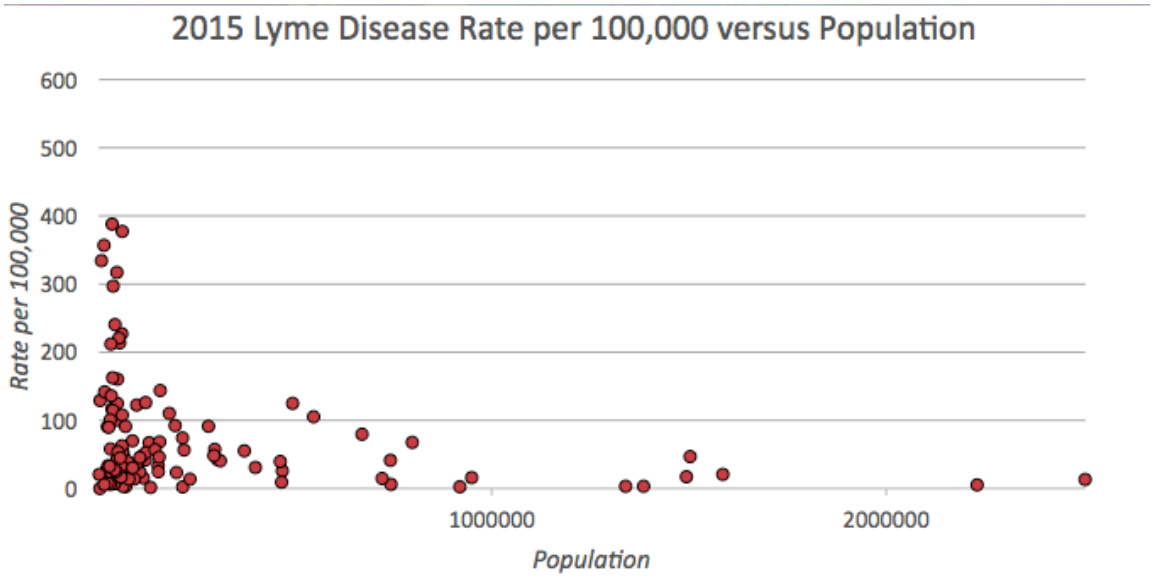

Figure 12. Scatterplot for 2015 Disease rate

The scatterplots displayed the change over the span of one year (the change between 2014 and 2015) and five years (the change between 2010 and 2015). The scatterplots present the association between population and the Lyme disease rate (Cases per 100,000 county population). From 2010 to 2015 there was a definite increase in cases overall, with a more dispersion in the lower left corner. From 2014 to 2015, there was one county with a notable decrease, but overall there was more cases in 2015.

Based on the above scatterplots, many counties contain a Lyme disease rate of less than 50 people per 100,000 population. In the scope of this study, the small number problem is not an issue for analysis using an OLS regression or a GWR (Meredith Franklin, personal communication). The small numbers (of county population) will not affect the regression results with the analysis methods chosen, as the OLS can handle the counties with 0 cases or a small number of cases.

4.2. Spatiotemporal Data Analysis for Lyme Disease

The spatiotemporal data analysis for Lyme disease was the first step in understanding the data of Lyme disease and its spatial and temporal trends. The following sections cover spatiotemporal trend of Lyme disease, emerging hotspots and local outliers.

4.2.1. Spatiotemporal trend of Lyme Disease

The results of the spatiotemporal analysis showed the overall growth of Lyme disease from 2000 to 2015. Figure 13 shows the spatiotemporal trend of the Lyme disease rate. The overall temporal trend was an increasing incidence of Lyme disease for the majority of counties in the study area. The only significant opposite trend lay on the southeastern portion of New York. This part of the New York state had a decreasing trend in the Lyme disease rate. The counties with the decreasing trend were Columbia County, Dutchess County, Nassau County, Orange County, Putnam County, Suffolk County, and Westchester County.

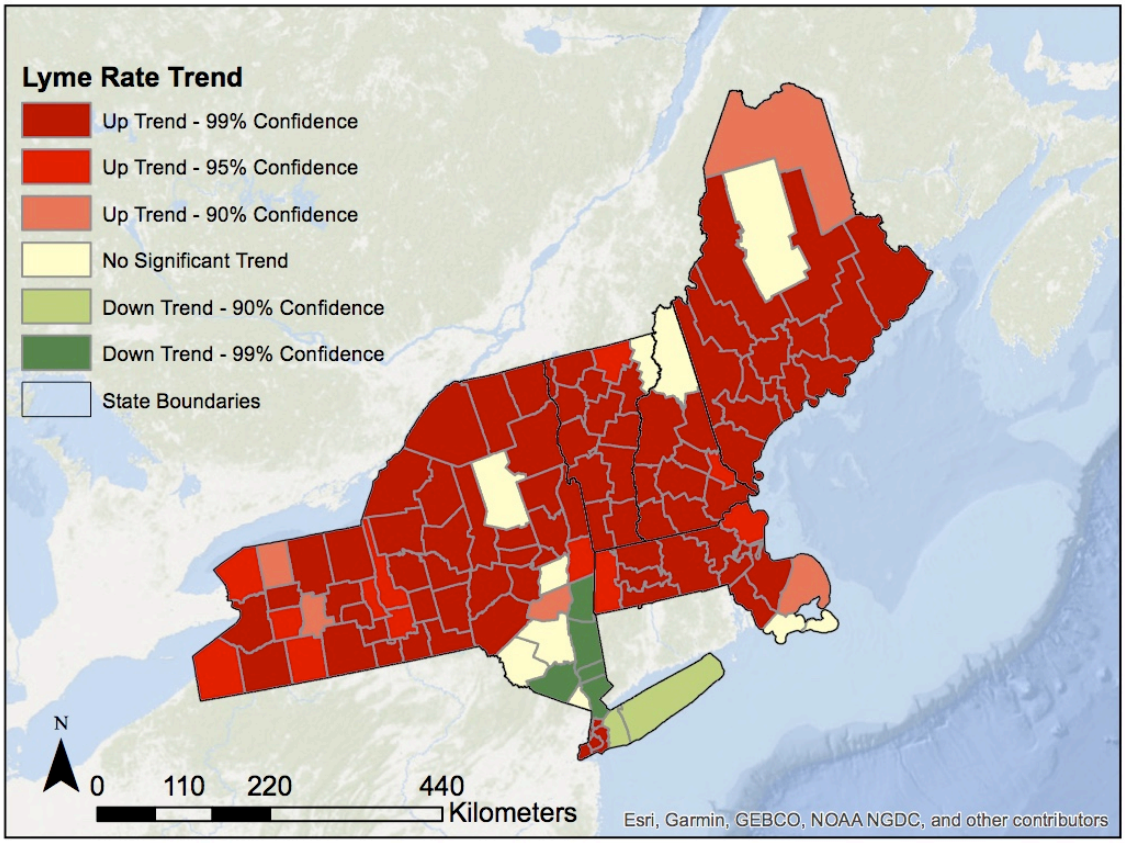

Figure 13. Overall trend of the Lyme disease rate per 100,000 population per county

4.2.2. Lyme Disease Hot Spots

This subsection demonstrates the locations of the Lyme disease hot spots. The Emerging Hot Spot Analysis looked at the data trend for each county, and established the stability and strength of the trend of the time steps to determine if the county has a statistically significant hot or cold trend. Figure 14 displays the emerging hot spot map for the Lyme disease count. Of the 116 counties, 90% of the counties (104 counties) were marked as no statistically significant patterns detected. There was only one emerging hot spot area identified, which comprised of both consecutive and sporadic hot spots. Recall that consecutive hot spots had two or more years in the start of the time period that were not significant, and every year after that was classified as significant. For a county to be classified as a sporadic hot spot, no cold spots could be present for

any years in those counties, and two or more years were not classified as statistically significant. Refer to Table 3 in Section 3.3.3. for the complete definitions of all the Emerging Hot Spot classifications in the study. Of the counties marked as a consecutive hot spot for Lyme disease, there was one county in New Hampshire (Hillsborough County) and six counties in Massachusetts (Essex County, Middlesex County, Worcester County, Norfolk County, Suffolk County, and Dukes County). Five counties marked as a sporadic hot spot were all located in Massachusetts (Hampshire County, Hampden County, Bristol County, Plymouth County, and Barnstable County).

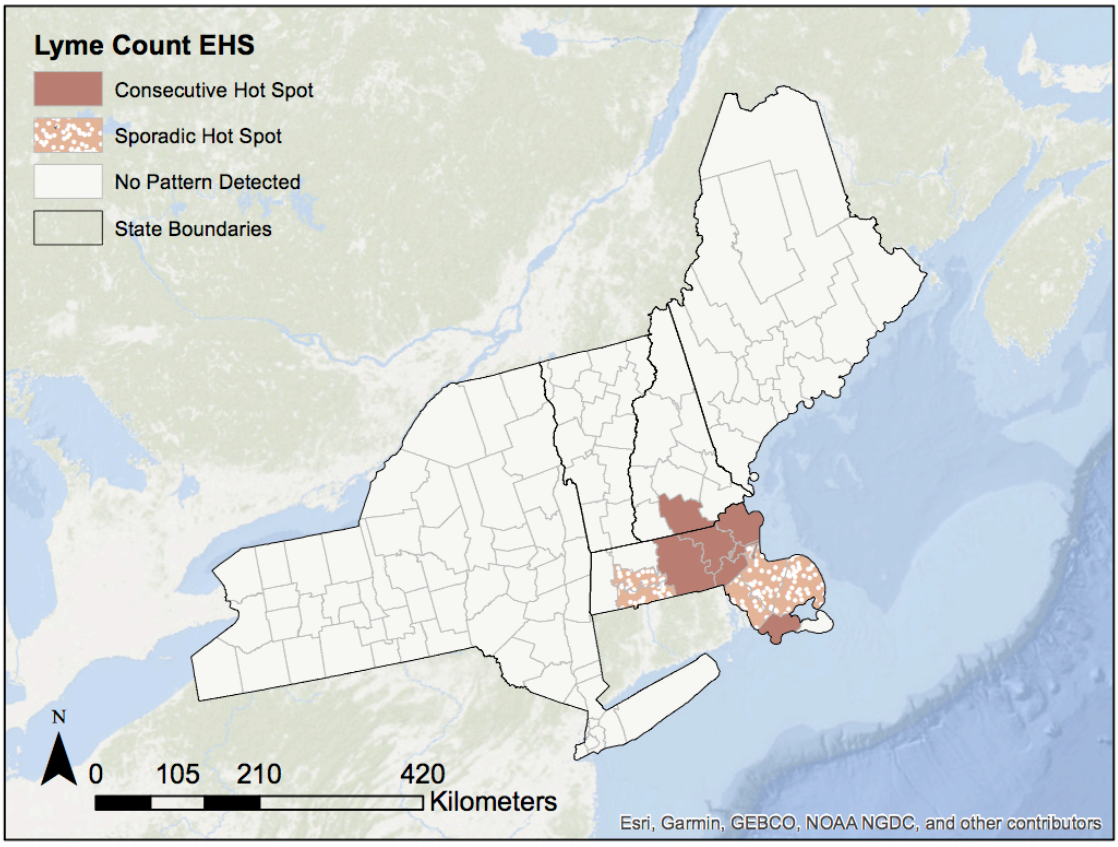

Figure 14. Emerging Hot Spots for Lyme disease count

When using the Lyme disease case count as the variable, the spatial and temporal trends were mostly unseen (Figure 14). This was due to a multitude of the counties with small

population counts which tend to have small numbers of confirmed cases. By using the incidence rate (cases per 100,000 population), the accurate risk or probability of contracting among the counties was compared. The emerging hot spot for Lyme disease rate was then created to correct this problem (Figure 15). While the Lyme disease rate reveals more variation in the data, no cold spots were found for Lyme disease in either Figure 14 or Figure 15. Definitions of the different types of Hot Spots can be found to Table 3 in section 3.3.3.

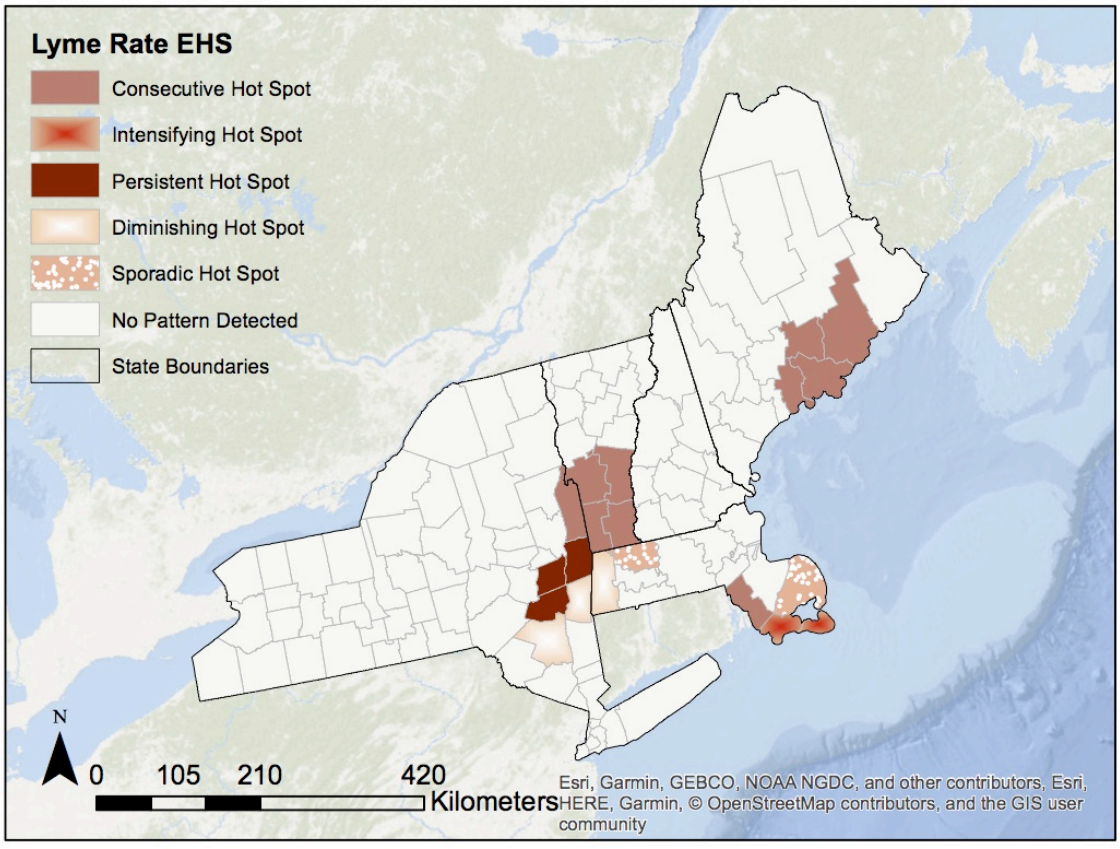

Figure 15. Emerging Hot Spots for Lyme disease rate per 100,000

The emerging hot spot map of Lyme disease rate per 100,000 population per county revealed more spatial and temporal trends (Figure 15). Not only were more variations in the types of emerging hot spots shown, but also more areas of emerging hot spots were found as well. Overall, three main areas were classified as emerging hot spot regions: The first hot spot

area was centrally located in the study area, consisting of 12 counties across the states of New York, Vermont, and Maine. These counties were Albany County, Columbia County, Greene County, Rensselaer County, Ulster County, Washington County, Bennington County, Rutland County, Windham County, Windsor County, Berkshire County, and Franklin County. There were five counties classified as consecutive hot spots, three counties classified as persistent hot spots, three counties classified as diminishing hot spots and one county classified as a sporadic hot spot. The persistent hot spot means the counties had been statistically significant for 90% of the years, and have no increasing or decreasing trend, while the diminishing hot spot means the counties were classified as statistically significant for 90% of the years, and have a declining intensity trend.

The spread of the second and third hot spot sites were confined to the eastern seaboard. The second emerging hot spot area consisted of four counties located on the southeastern coastal counties in Maine. The counties in this hot spot were Hancock County, Knox County, Lincoln County, and Waldo County. All four counties were classified as consecutive hot spot. The final emerging hot spot area was located in the southern tip of Massachusetts. The four counties in this hot spot site were Barnstable County (a sporadic hot spot), Bristol County (a consecutive hot spot), Dukes County (an intensifying hot spot), and Nantucket County (an intensifying hot spot). The intensifying hot spot in the disease rate map shows the counties were statistically significant for most of the years, and have an increasing trend in cases. The location breakdown of the emerging hot spots can be seen in Appendix A.

When the Lyme disease rate map (Figure 15) was compared to the Lyme disease count map (Figure 14), the number of the study area classified as hot spots detected increased by eight counties (approximately 7%). There was an increase of three consecutive hot spots (10 total), a

decrease of three sporadic hot spots (2 total), and the addition of three persistent hot spot counties, three diminishing counties, and two intensifying counties. The hot spot in Massachusetts decrease in the Lyme disease rate map, but there were hot spots on the eastern coast of Maine, and the intersection of New York, Vermont and western Massachusetts. Overall the Lyme disease rate showed a more evident trend than the Lyme disease count.

There was some association between the Emerging Hot Spot map for the Lyme disease rate (Figure 15), and the overall Lyme disease trend map (Figure 13). The location of the diminishing hot spot in the Lyme disease incidence map correlated with the downward trend of the rate trend map. Because those counties are reporting less cases each year, but still higher than the rest of the map, it is classified as a diminishing hot spot. Despite the scarcity of hotspots found in the hot spot analysis, the overall trend of Lyme disease cases in each county was increasing over time.

The comparison of the Lyme disease trend map (Figure 14) and the Lyme disease count map (Figure 13) provided little extra information. The hot spot for Lyme disease count was almost completely classified as upward trend. The one exception is Dukes County on the southeastern portion of Massachusetts. This is the only county out of the 12 in the hot spot that was classified as having no trend.

4.2.3. Local Outlier Analysis of Lyme Disease

The Local Outlier analysis was included as a supplementary analysis for the spatiotemporal trend of Lyme disease in Section 3.3.5. Figure 16 shows spatial clusters and outliers of the Lyme disease case count in the study area. There were 71 counties (62%) categorized as Only Low-Low Cluster, meaning counties with low disease counts surrounded the county with low disease counts. The spread of the Low-Low cluster paralleled the dispersion of

the no pattern detected category in the Lyme disease count. The Only High-High cluster on the local outlier map followed a similar pattern as the consecutive and sporadic hot spot in the emerging hot spot map. There were a total of 21 counties classified as High-High Cluster. The nine counties in the Multiple Types category were intermittently scattered throughout the study area. They were noted as having no spatiotemporal trend detected in the Emerging Hot Spot Analysis. The four counties marked as Only Low-High Outlier generally corresponded to the counties marked as no pattern detected except Dukes County, Massachusetts, which was marked as a consecutive hot spot. The other three Low-High Outlier counties, Cheshire County in Vermont, Franklin County in Massachusetts, and Bronx County in New York were all located next to a hot spot. The definitions and explanations of the different categories are in Table 4.

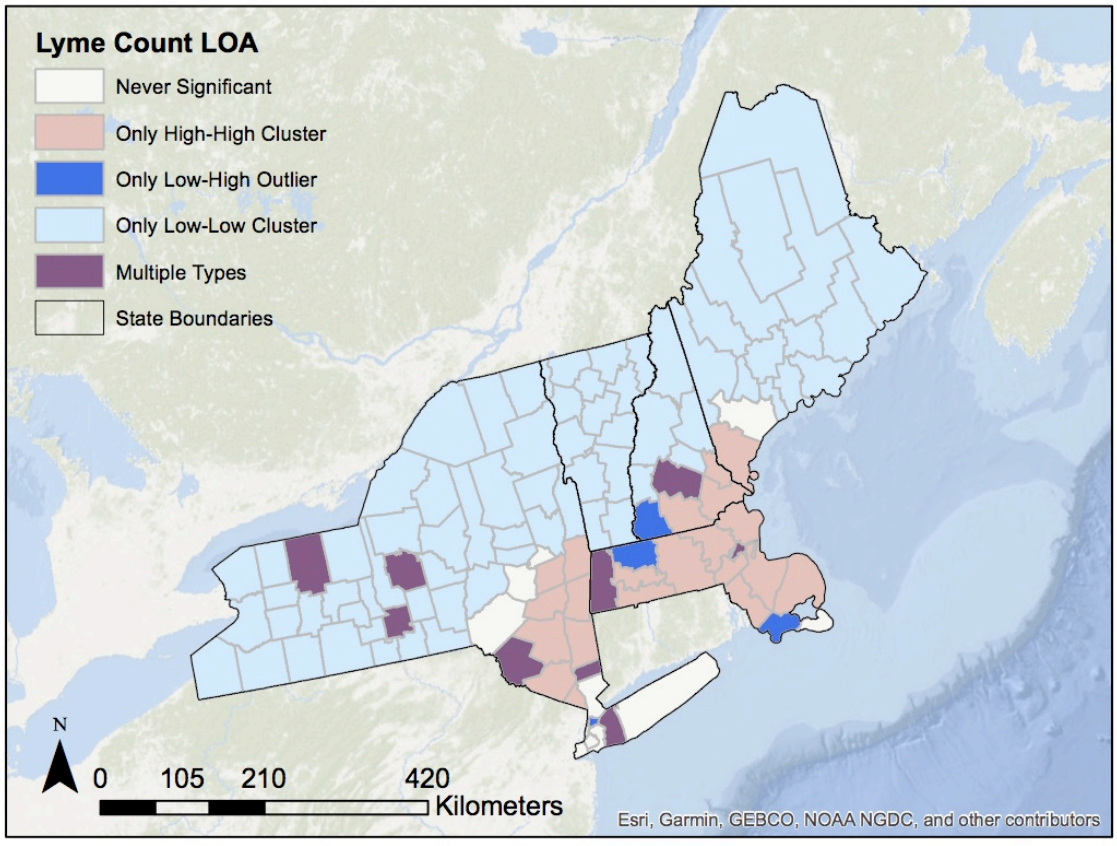

Figure 16. Local spatial clusters and outliers for Lyme disease cases

Surprisingly, the Local Outlier Analysis map for the Lyme disease rate has a different distribution of values (Figure 17). The pattern of the 13 Multiple Types counties paralleled the outline of the consecutive hot spot fairly well, with a few extra counties identified in this category. As with the Lyme disease rate, the Only Low-Low Cluster tracked the outline of the no detected pattern category. There were 80 counties (69% of the total counties) under this classification. Eleven counties were classified as Only High-High Clusters, which followed the other hot spot locations similarly. One county (Delaware County in New York) classified as a Low-High Outlier was located next to a few High-High Clusters, and was detected as a hot spot in the hot spot map.

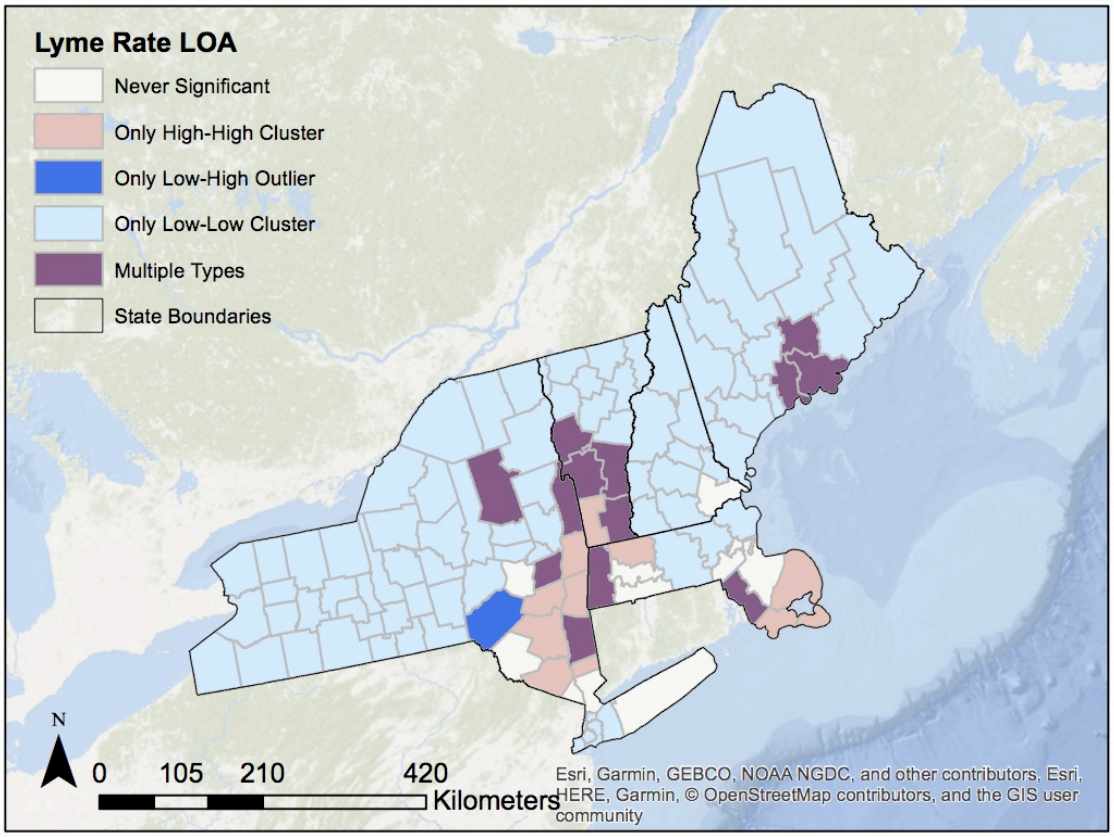

Figure 17. Local spatial clusters and outliers for Lyme Disease Rate

There were minor differences between the Local Outlier maps for the count and rate of Lyme disease. In the Local Outlier map for Lyme disease rate (Figure 17), a smaller number of counties were marked as High-High Cluster in Massachusetts, which were classified as High-High Clusters in the Local Outlier map for the disease count (Figure 16). The other major difference was the increase of nine counties, or 8% of the number of counties, marked as Low-Low Cluster in the Local Outlier map of the rate compared to that of count.

4.3. Spatiotemporal Data Analysis for Environmental Factors

The following subsections describe the result of the spatiotemporal data analysis for the environmental factors. The results for climate characteristics are included in Section 4.2.1 (for precipitation) and Section 4.2.2 (for temperature). The result of the Emerging Hot Spot is shown, followed by the Local Moran's I result. Section 4.2.3 contains optimized hot spot map for the forest cover data. At the end of each section is a comparison of the Lyme disease maps and each environmental factor. Any correlation between the maps was not taken as the cause of the Lyme disease hot spots; rather, the correlations were noted to evaluate if they have a positive or negative relationship. The basis of this process was to assess if there was any temporal trend through the data that was occurring in the same locations as the Lyme disease hot spots.

4.3.1. Precipitation

The annual precipitation data had a definite cold temporal trend during 2001 to 2015. Figure 18 shows the Emerging Hot Spots for precipitation. Interestingly, there were no temporal hot spots. In the western portion of New York, there was a decrease in precipitation over the time period. Several counties in this region (Niagara County, Orleans County, Monroe County, Genesee County, Livingston County, Ontario County, Yates County, Seneca County, Schuyler County, and Steuben County) were persistent cold spots that have been a statistically significant

cold spot with no increasing or decreasing trend for 90% of the time period. Several counties with decreasing precipitation trend surrounded the persistent cold spots. These sporadic cold spot included nine counties (Wyoming County, Allegany County, Jefferson County, Wayne County, Cayuga County, Cortland County, Tompkins County, Tioga County, and Chemung County). Besides western New York, three neighboring counties in Vermont (Grand Isle County, Chittenden County, and Washington County) were sporadic cold spots and two counties in east-central New York (Schoharie County and Schenectady County) were oscillating cold spots, meaning there were decreasing precipitation trends in the final time period with a historical hot spot at one point during 2001 and 2015. No statistical significant trends (hot spots or cold spots) were detected for the rest of the study areas.

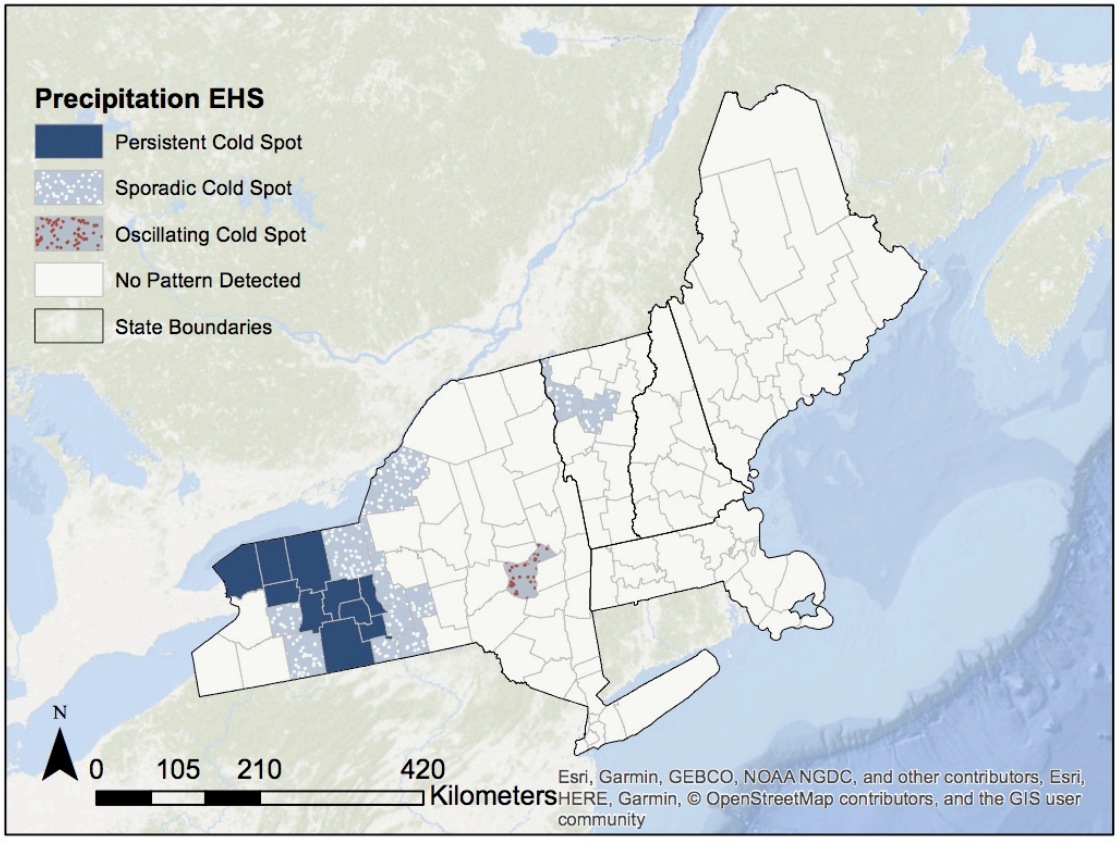

Figure 18. Emerging Hot Spots for Precipitation

The local outlier analysis revealed in Figure 19 provided interesting insight into the precipitation trend. The majority of the map, 68 counties, was classified as multiple types, which shows the precipitation for those counties varied over the years. There was a distinct linear trend of the spread of this value through the study area. There was a collection of five counties in central New York that are classified as high-high cluster (Madison County, Otsego County, Oneida County, Herkimer County, and Fulton County). There was also one county in Vermont (Bennington County) and seven neighboring counties in Massachusetts (Essex County, Norfolk County, Bristol County, Middlesex County, Suffolk County, Plymouth County, and Barnstable County). These areas show where there was a statistically significant cluster of high values of precipitation.

There were seven counties total that have No Significant trend. There were five scattered throughout New York (Chautauqua County, Onondaga County, Lewis County, Franklin County, and Suffolk County), and two neighboring in Massachusetts (Dukes County and Nantucket County). There were three total outlier counties, two High-Low outliers and one Low-High outlier. They were all located in New York. The High-Low outlier counties were Chenango County and Essex County, while the Low-High outlier is Richmond County. These show the neighboring values were conflicting with the precipitation values for the county.

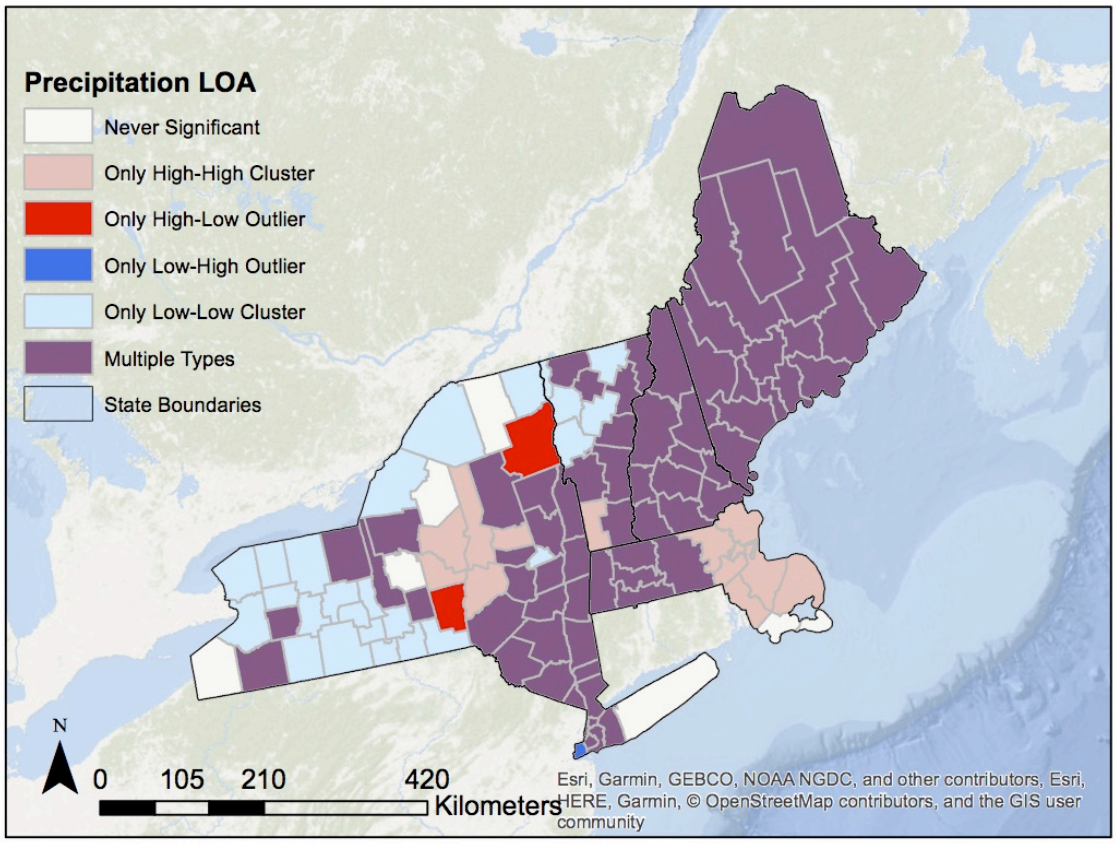

Figure 19. Local spatial clusters and outliers for Precipitation

Compared to emerging hot spots (Figure 18), spatial clustering by the Local Outlier Analysis (Figure 19), yielded more local variations. The large cold spot in western New York was almost exclusively marked as a Low-Low Cluster. A few of the sporadic cold spot counties were marked as Multiple Types. This meant that they have a combination of trends. There were two Only High-Low Outliers in New York (Chenango County and Essex County), and thirteen High-High Clusters scattered in New York, Vermont, and Southern Massachusetts. Surprisingly, there were no statistically significant hot spots that share the same locations. Other than a small county in New York (Richmond County) marked as Low-High Cluster, the rest of the study area was marked as Multiple Types. This extreme variation was intriguing when compared to the consistency in the Emerging Hot Spot map.

The trend of the precipitation Emerging Hot Spot map (Figure 18) did not overlap with those of the Lyme disease (Figures 14, and 15). This suggests that precipitation was not a strong predictor of Lyme disease.

4.3.2. Temperature

The overall spatiotemporal trends for the three temperature variables during 2000 and 2015 were similar. The emerging hot spots for maximum temperature can be seen in Figure 20, the emerging hot spots for mean temperature can be seen in Figure 21, and the emerging hot spots for minimum temperature can be seen in Figure 22. The similarities of these maps will be addressed first, followed by their differences. One remarkable trend is that Grand Isle County, a small county in Western Vermont on the border of New York, was always classified as having no pattern detected in all three temperature maps. This was remarkable as it is in the middle of a contiguous cold spot.

The northern and northeastern portions of Maine, New Hampshire, Vermont, and New York had a strong decreasing trend in temperature during 2000 and 2015. The majority of the counties in Maine and New Hampshire show persistent cold spots in the outputs of the emerging hot spot analysis. Surrounding the counties with persistent cold spots were those of sporadic cold spots throughout the four states.

In terms of increasing trends in temperature, there were two hot spots in the southeastern portions of Massachusetts and New York. These hot spots areas were either persistent hot spots (continuously increases in temperature) or sporadic hot spots (inconsistently classified as hot spots). In between the cold spots and the hot spots was a string of counties with no significant spatiotemporal patterns of temperature detected. Generally speaking, it the data showed that the weather has been getting colder in the northern section of the study area, and increasingly getting

warmer in the southern section. This followed the logic that the temperature decreases the further north traveled.

There were few differences between the maps of the emerging hot spots in the annual maximum temperature (Figure 20), the annual mean temperature (Figure 21), and the annual minimum temperature (Figure 22). The emerging hot spot map for maximum temperature showed sporadic hot spots throughout the most areas of Massachusetts, whereas those maps for mean temperature and minimum temperature were classified as persistent hot spots or intensifying hot spots. The sporadic hot spots of maximum temperature were also significantly larger in areas than those of mean and minimum temperature. The minimum temperature map had the only counties classified as a New Cold Spot. The two counties were Delaware County and Greene County, and were centrally located in New York. These minor differences in the maps below are why all of the temperature variables were included in the stepwise regression. The temperature is not consistent throughout the study area, thus the inclusion of multiple values allows for the possibility of one aspect of temperature correlating to Lyme disease more than another.

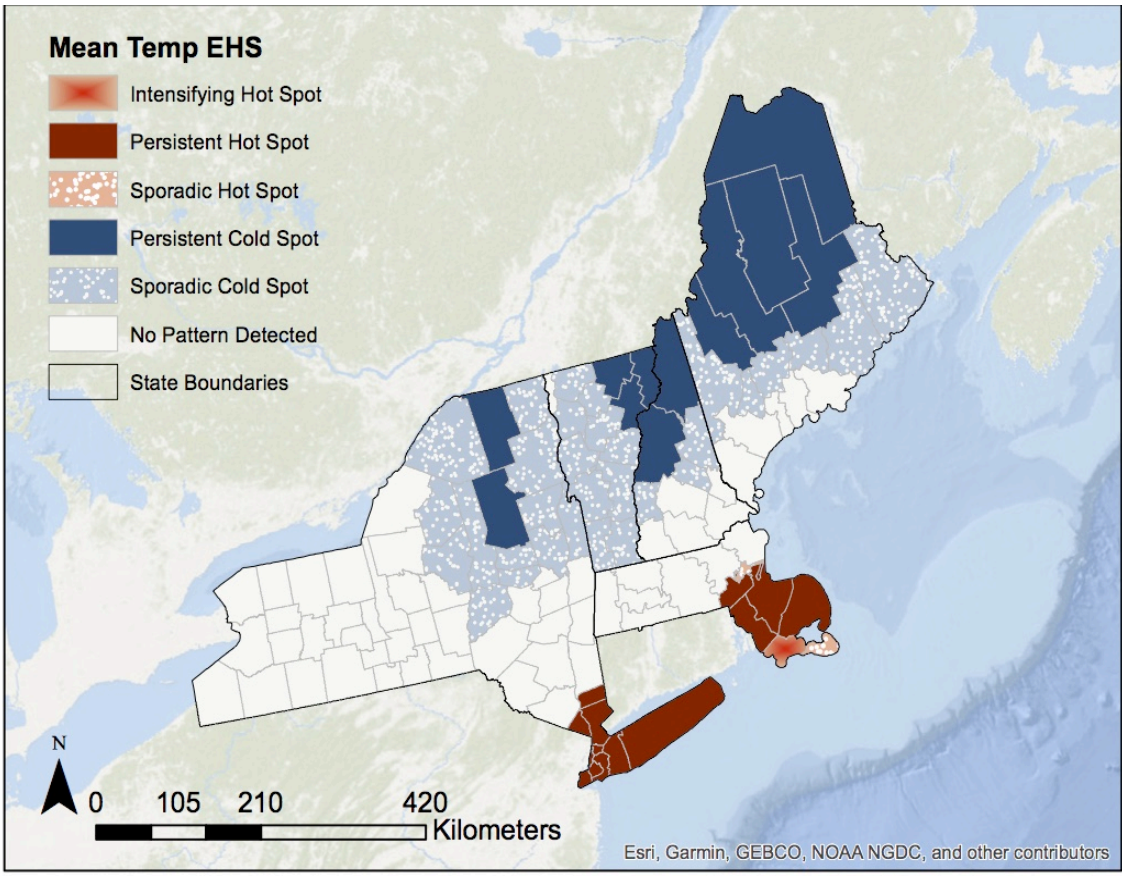

Figure 20. Emerging Hot Spots for Maximum Temperature

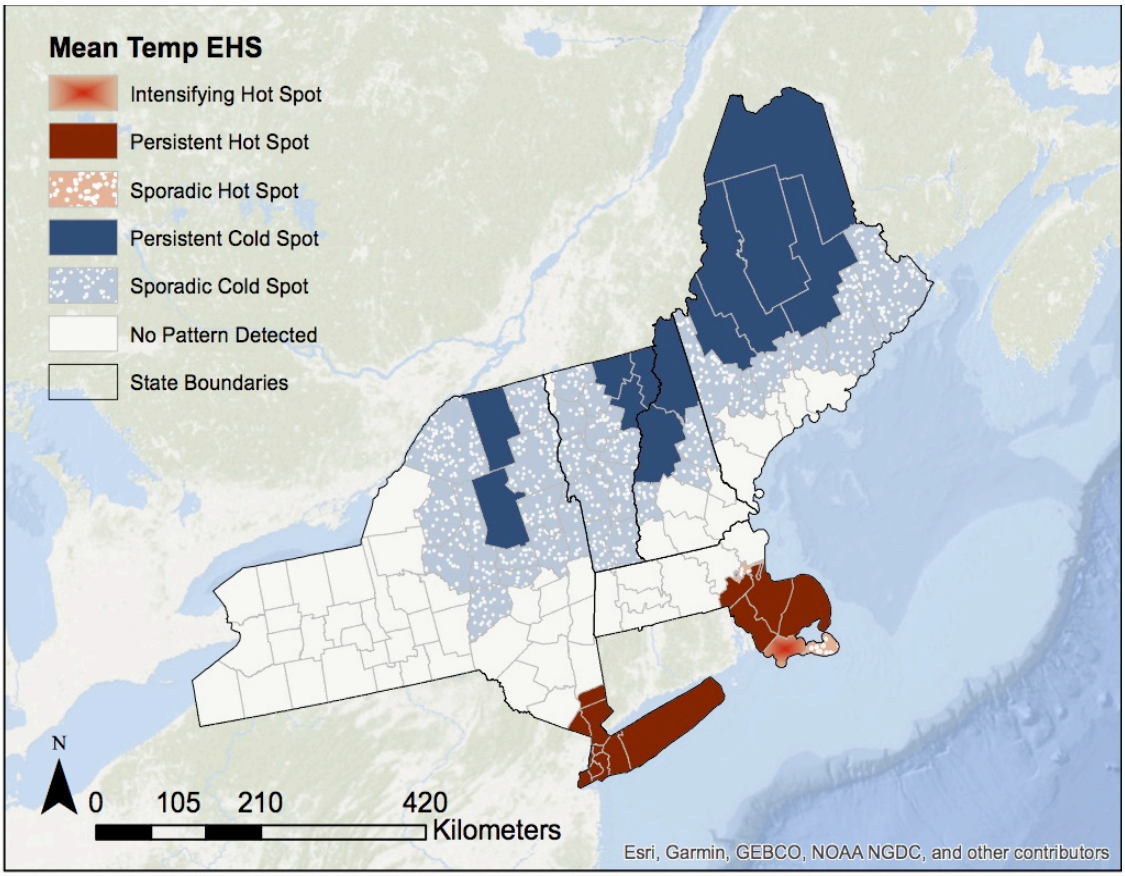

Figure 21. Emerging Hot Spots for Mean Temperature

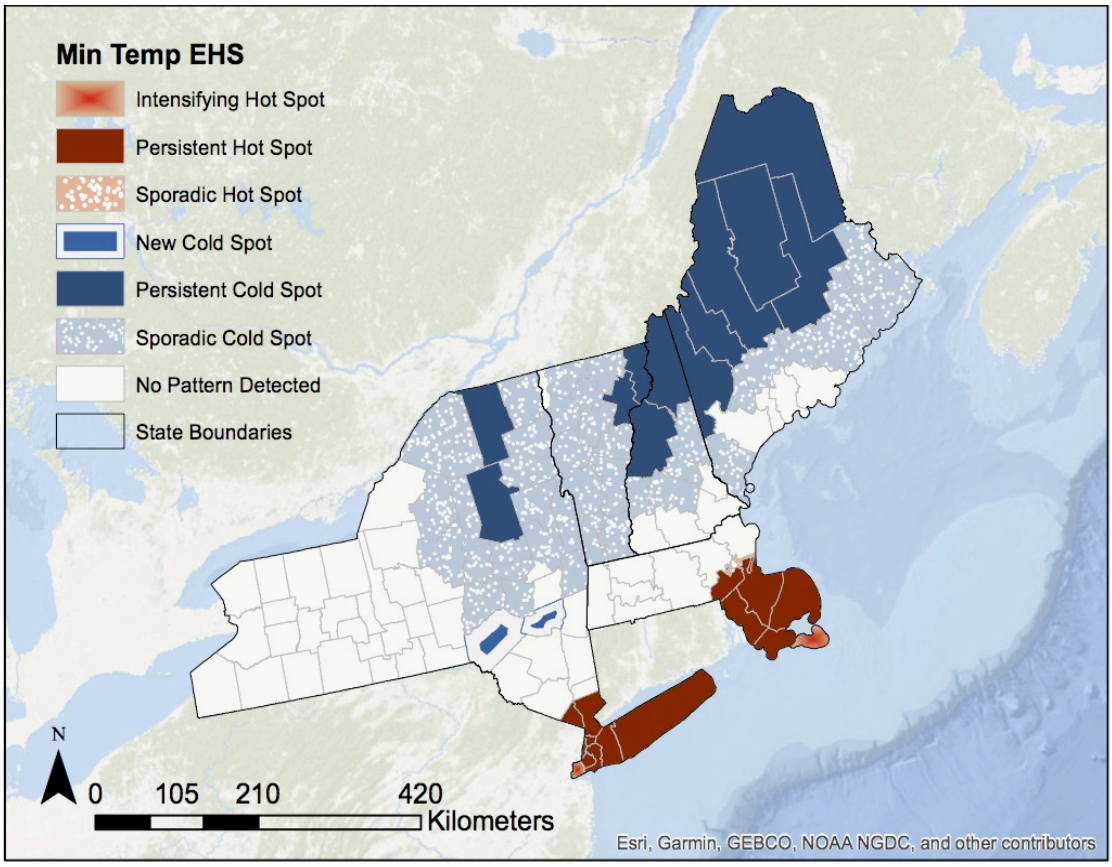

Figure 22. Emerging Hot Spots for Minimum Temperature

Visually, the temperature emerging hot spots (Figure 20, 21 and 22) exhibit similarities to the Lyme disease emerging hot spots (Figure 14 and 15). While the cold spots were not reflected in the Lyme disease maps, the hot spot located in Massachusetts was analogous to the hot spot depicted in the Lyme disease count map. The Lyme disease rate map had a small portion of the temperature hot spot replicated in Massachusetts (in Bristol County, Barnstable County, Dukes County, Nantucket County), while the rest of the Lyme disease hot spots fell under or near temperature cold spots. There were nine counties that were marked as cold spots, two in Maine (Hancock County and Waldo County), four in Vermont (Bennington County, Rutland County, Windham County, and Windsor County), and three in New York (Washington County, Rensselaer County, and Greene County). This result was the opposite of what was anticipated for

the relationship between temperature and Lyme disease rate. The expectation was based upon a negative effect of the colder weather to the tick life cycle lead to less cases of Lyme disease. Conversely, it was assumed warmer weather would increase cases of Lyme disease due to increased movement in tick populations, and the human population. This was reflected in the hot spot in southern Massachusetts, but not in the other Lyme disease hot spots. The lack of a clear correlation demonstrated there was a complex relationship between temperature and cases of Lyme disease.

On the other hand, the results of local outlier analysis (or local Moran's I) of the temperature variables alone showed more variation in spatial clustering (Figures 23, 24, and 25). There is a distinct Northeast to Southwest trend of the Low-Low Clusters. Scattered throughout there is a few multiple types counties, with three areas of High-High Clusters in the southeastern portion of Massachusetts, and the southern and western portions of New York. There is a barrier of counties marked as Never Significant in between the two types of clusters.

While the number of counties in the different categories was fairly equivalent in the local outlier maps, there are a few differences. The mean temperature map had the only county marked as High-Low Outlier, which was Franklin County, Massachusetts. It was marked as Never Significant in the maximum temperature, and is a part of the Low-Low Cluster for minimum temperature. The maximum temperature local Moran's I map had the only Low-High Outlier, in Greene County, New York. Greene County was marked as Never Significant in both the minimum and mean maps. The final difference is the minimum temperature map had more Never Significant and Low- Low Cluster, and less Multiple Types and High-High Cluster counties than the other two maps.

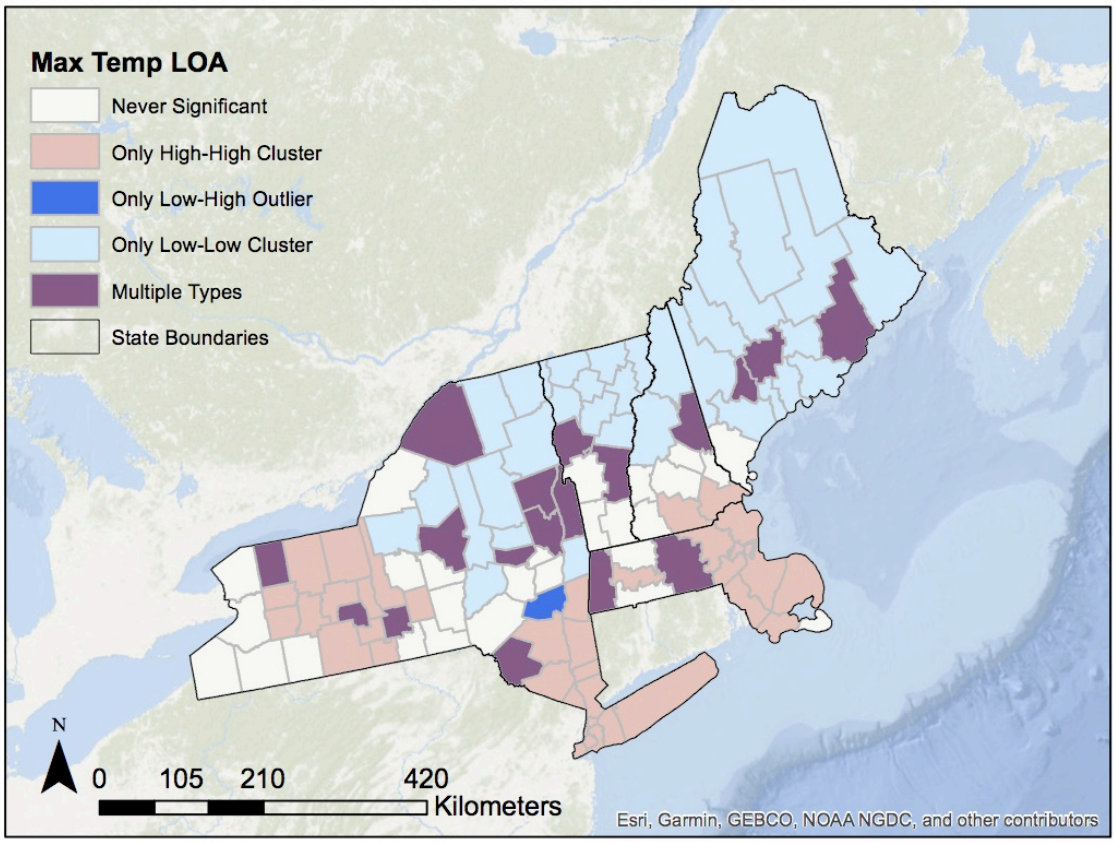

Figure 23. Local spatial clusters and outliers for Maximum Temperature

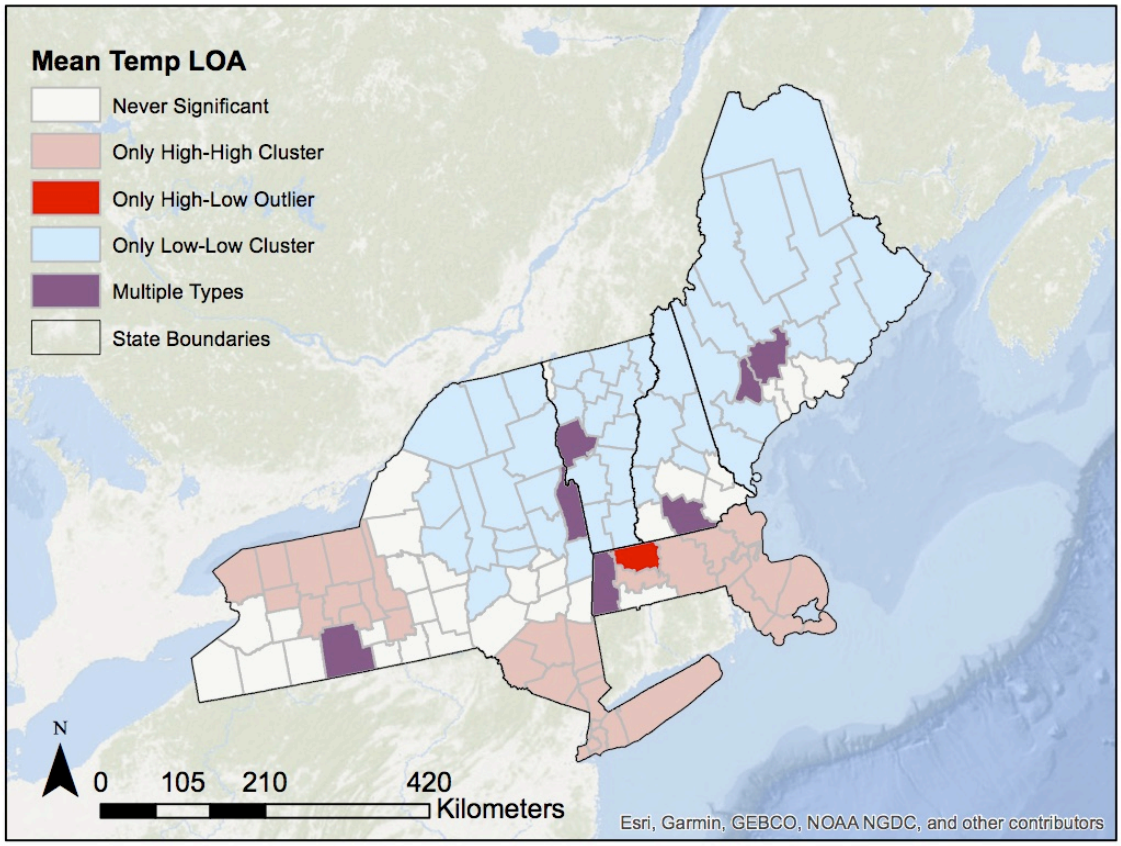

Figure 24. Local spatial clusters and outliers for Mean Temperature

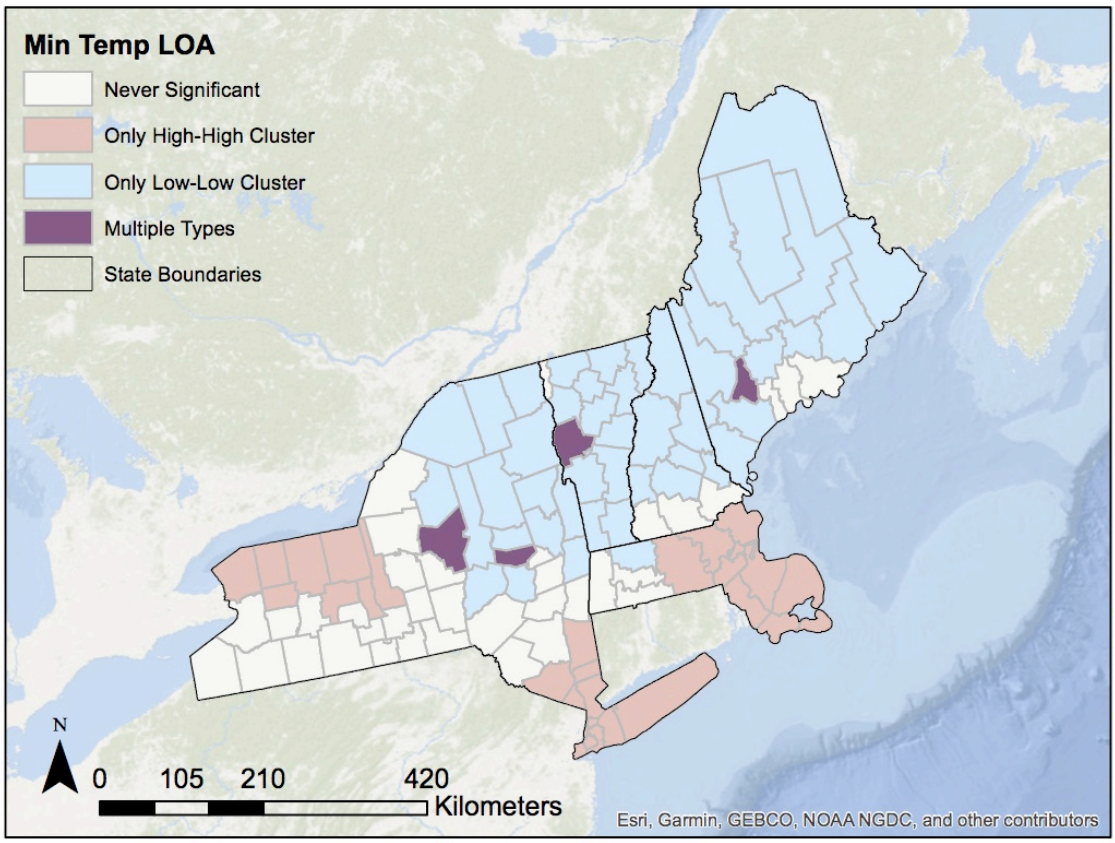

Figure 25. Local spatial clusters and outliers for Minimum Temperature

When compared to the temperature Emerging Hot Spot maps (Figures 20, 21, and 22), the trend of these outlier maps (Figures 23, 24, and 25) was comparable. The Northeastern pattern of the cold spots in the Emerging Hot Spot maps was almost exclusively Low-Low Clusters. The Southeastern portions of Massachusetts and New York had mostly High-High Clusters, with another High-High Cluster in western New York, which was similar to the hot spots found in the maps.

The Local Outlier Analysis maps (Figures 23, 24, and 25), of the temperature variables were similar in arrangement to the Lyme disease rate and count maps (Figures 14 and 15). Temperature seemed to have a strikingly similar trend to Lyme disease counts, and less of a correlation with Lyme disease rate. However, there were still similarities with the Lyme disease

rate maps. Lyme disease rate Local Outlier Analysis map had smaller groups of High-High Clusters, and more overall Low-Low Clusters. Lyme disease case count overlap closely with the mean temperature and minimum temperature Local Outlier maps. The main differences were a few Low-High Clusters, and less Never Significant areas in the Lyme case counts. It is possible temperature was a better predictor in true case counts as opposed to the incidence rate of Lyme disease. However, the Lyme disease rate is a more robust analysis variable.

4.3.3. Spatial Data Analysis Results for Forest Cover

The Optimized Hot Spot Analysis tool was used for the forest cover variable, as there was only one year of data available for each county (Figure 26). A sizeable contiguous hot spot area, comprised of 63 counties, was found from the eastern New York border all the way to Maine. Several cold spot areas were found in the southern and western portions of New York, the southeastern portion of Massachusetts, and a section on the eastern Maine coast. These areas consisted of 42 counties, or 36% of the study area. Scattered intermittently throughout were 11 counties that were not statistically significant for either hot or cold spots.

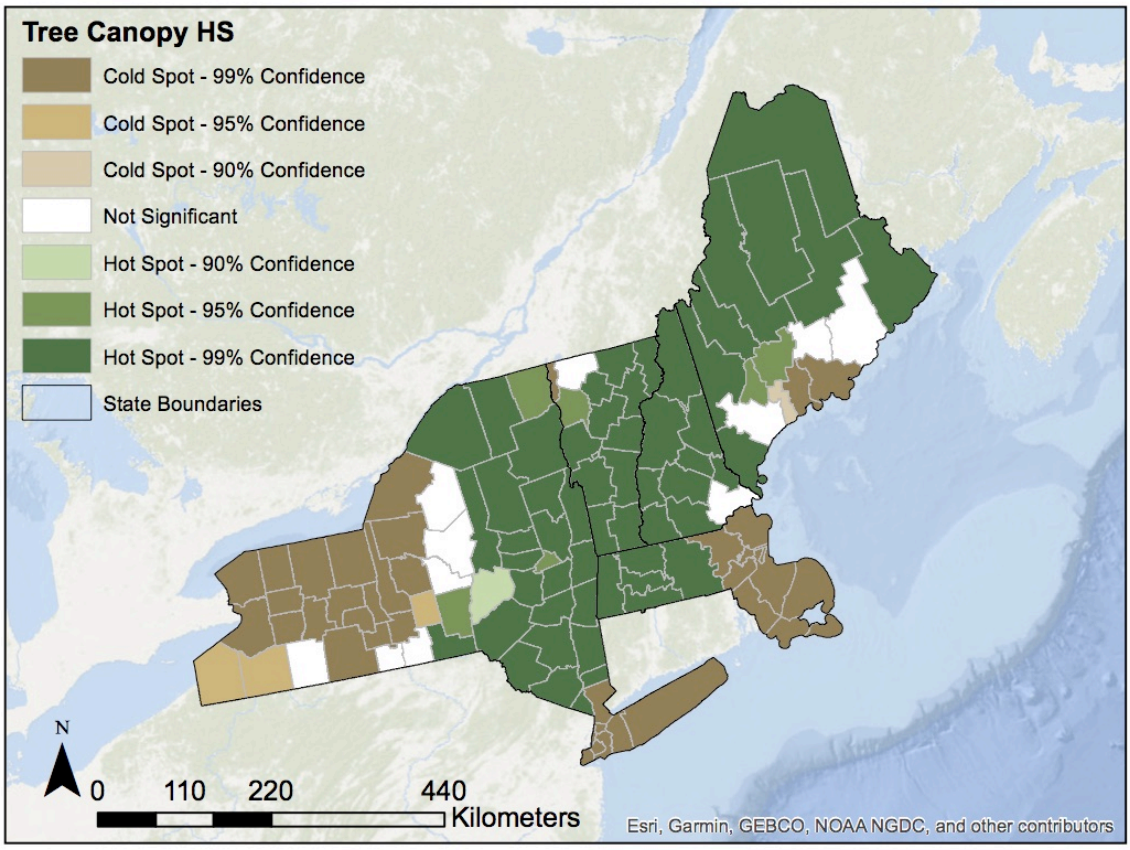

Figure 26. Optimized Hot Spots for Forest Cover Coverage

The visual comparison of the Lyme disease hot spots map (Figure 14 and 15) and the forest cover hot spot map (Figure 26) indicated an important correlation between the two variables. The six of the counties marked as hot spots in the Lyme disease maps corresponded to cold spots in the forest cover map. The counties that were classified as cold spots were Knox County, Lincoln County, Barnstable County, Bristol County, Dukes County, and Nantucket County. The first two counties were located in eastern Maine, and the last four in southeastern Massachusetts. Of the fourteen other counties marked as hot spots in the Lyme disease rate map, two counties in Maine were marked as not significant, and twelve were marked as hot spots in the forest cover map. The original relationship for areas with higher forest cover was a higher correlation to areas with more cases of Lyme disease. However, it appears that the relationship

was not as clear as expected. This could be due to an increased population in the areas with lower percentages of forest cover.

4.4. Lyme Disease Models

The final segment of the project results included the model selection. The OLS regression results and the final regression model were described. Selected models are demonstrated in section 4.4.1. For the complete list of models, see Appendix B. A residual analysis result was conducted to present how well the final model fit.

4.4.1. Stepwise Regression Model for Lyme Disease Rates

The methods chapter described the backwards stepwise regression that was completely run twice. After the first iteration, it was determined that more variables were needed to improve the accuracy of the model. The second iteration included the additional variables. From the stepwise regressions, there were five regressions determined to fit the data best:

- OLS 3 Model: Precipitation & Max Temperature
- OLS 5 Model: Year, Longitude, Latitude, Precipitation, Mean Temperature, Maximum Temperature, Minimum Temperature, Forest Cover
- OLS 9 Model: Year, Longitude, Latitude, Mean Temperature, Max Temperature, Forest Cover
- OLS 12 Model: Year, Longitude, Latitude, Max Temperature
- OLS 13 Model: Year, Longitude, Latitude, Mean Temperature

These regressions are shown in Table 6.

Table 6. Summary of test results selected models

	OLS 3	OLS 5	OLS 9	OLS 12	OLS 13
Dependent Variables	Precipitation, Maximum Temperature	Year, Longitude, Latitude, Precipitation, Mean Temperature, Maximum Temperature, Minimum Temperature, Forest Cover	Year, Longitude, Latitude, Mean Temperature, Maximum Temperature, Forest Cover	Year, Longitude, Latitude, Maximum Temperature	Year, Longitude, Latitude, Mean Temperature
Variance Inflation Factor> 7.5	Yes	No	No	Yes	Yes
Statistically Significant Coefficients	Yes	No	Yes	Yes	Yes
Jarque Bera not Significant	Yes	Yes	Yes	Yes	Yes
Koenker Test Significant	No	No	No	No	No
Residuals Correlated	Yes	Yes	Yes	Yes	Yes
AICc	22544.586	22310.362	22308.602	22328.280	22312.169
Adjusted R-squared	0.029	0.1462	0.1465	0.1365	0.1440

With the exception of the OLS 5 model, the partial coefficients for all of the independent variables had the anticipated signs (+ or -) and were statistically significant. The Variance Inflation Factor (VIF) of the OLS 5 and OLS 9 models are significantly over 7.5, indicating the considerable variable redundancy (or collinearity). Once those variables were removed, the values of the VIF became acceptable (< 7.5) in the OLS12 and OLS 13 models. All five models passed the Jarque Bera test and the Koenker Test, but the residuals were still spatially autocorrelated. This indicates no model bias and the stationary relationships between the

explanatory variables and the Lyme disease rate, but there is likely regression misspecification, meaning some key explanatory variables are missing. Stationarity refers to the consistent relationship with Lyme disease in reference to geography and time period.

The presented regression models were chosen to show the increased capability of the regression model by adding more variables. OLS 3 was the result from the first complete stepwise regression and showed that only approximately 3% of the Lyme disease rate was explained by the regression. While OLS 5 did not pass the majority of the model checks, it was included to show the substantial increase in the Adjusted R-squared, and the decrease in the AICc from OLS 3. In OLS 5 and OLS 9, there was a high VIF due to the multicollinearity between the mean temperature and maximum temperature. The VIF check was passed when only one temperature variable was included in the model. This is shown in OLS 12 and OLS 13.

The final model choice was between OLS 12 and OLS 13. The results were sufficiently comparable that either model could have been selected. There was an improvement of than 1% in the Adjusted R-squared of OLS 13, which was selected as the ultimate regression model. This model was picked as the overall best fit with the variables available.

The regression results were different than expected. Precipitation and forest cover had a smaller effect than anticipated. The anticipated relationship with higher precipitation leading to more cases of Lyme disease (due to increased host and habitat survivability) was not found. The location (longitude and latitude) and time (year) were consistently significant, showing spatial locations and time were noteworthy predictors to Lyme disease. Temperature was a significant predictor itself, but the inclusion of other variables determined which aspect of temperature was significant in the different regressions. Overall, the regression models in this study did not explain a sizeable portion of the variance of the Lyme disease rate. Other (natural or human)

environmental variables might increase model performance. Other types of modeling techniques (e.g. Poisson regression) should also be considered for Lyme disease model prediction.

The OLS 13 model was selected to be the final Lyme disease model. The full OLS 13 model equation is as follows:

$$y = -4057.99 + 3.56X_1 + 12.42X_2 - 42.05X_3 - 6.69X_4 + 0 \quad \text{eq. 8}$$

4.4.2. Residual Analysis Results for the Lyme Disease Models

The residual analysis was included to evaluate the model performance. This consisted of statistics of the residuals, a residual plot, and a residual map. Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Bias Error (MBE) were calculated for both OLS 12 and OLS 13 (Table 7). The results displayed that OLS 13 was overall the model with the least errors and thus the best choice of the Lyme disease model.

Table 7. Residual analysis results

Models / Residual Analysis	**OLS 12**	**OLS 13**
RMSE	98.79	98.36
MAE	51.10	52.23
MBE	-3.36	-2.59

The RMSE, as a measure of accuracy determined which model with the smallest error for the dataset. When comparing OLS 12 and OLS 13, the latter had a marginally smaller RMSE. To determine the best overall model, the next two residual statistics were then examined. The comparison of the two models with MAE, a measure of the average absolute difference between the observed values and the predicted values, showed a slightly smaller MAE of OLS 12 than that of OLS 13.

The final residual statistic MBE for the overall bias of the model was then evaluated. It is important to note that the positive and negative difference could cancel out in MBE, displaying whether the model uniformly over- and under-estimated. Both models had a negative MBE, indicating that the model over-estimated the Lyme disease rate more than it under-estimated the rate. The OLS 13 model had MBE closer to 0, which was the ideal balance between the counties with over-estimated Lyme disease rate and those with under-estimated rate. This supports the previous statement that the OLS 13 model was the better fit than OLS 12. However, the differences between the two models were marginal, thus choosing OLS 13 over OLS 12 would not result in a pronounced improvement.

The residuals were also plotted against the predicted values (Figure 27). This plot shows that the model can be improved based on the dispersion of the residuals. The residual plot is not evenly distributed vertically, they have clear outliers, and they have a distinct linear trend. The plot is showing heteroscedasticity, meaning the residuals are getting larger as the prediction values increase. This indicates there is another variable influencing the incidence of Lyme disease that is not included in the regression equation.

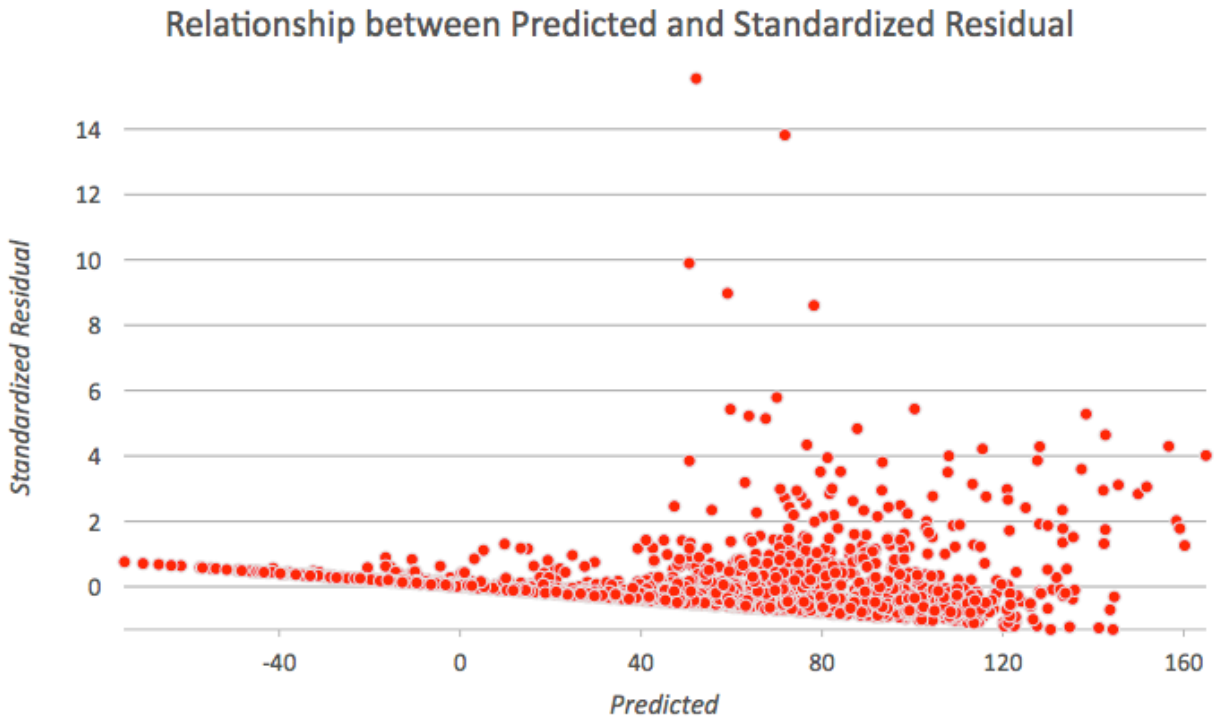

Figure 27 Residual plot

The final step of the residual analysis was to create a residual map (Figure 28). This was created to show the disparity in the predicted versus actual values of Lyme disease in OLS 13. The areas in blue show the locations where the model over-predicted the Lyme disease rate. Conversely, the areas in red show the locations where the model under-predicted the Lyme disease rate. Overall, there were 25 counties (or 22%) marked as under-predicted, and 43 counties (or 37%) marked as over-predicted.

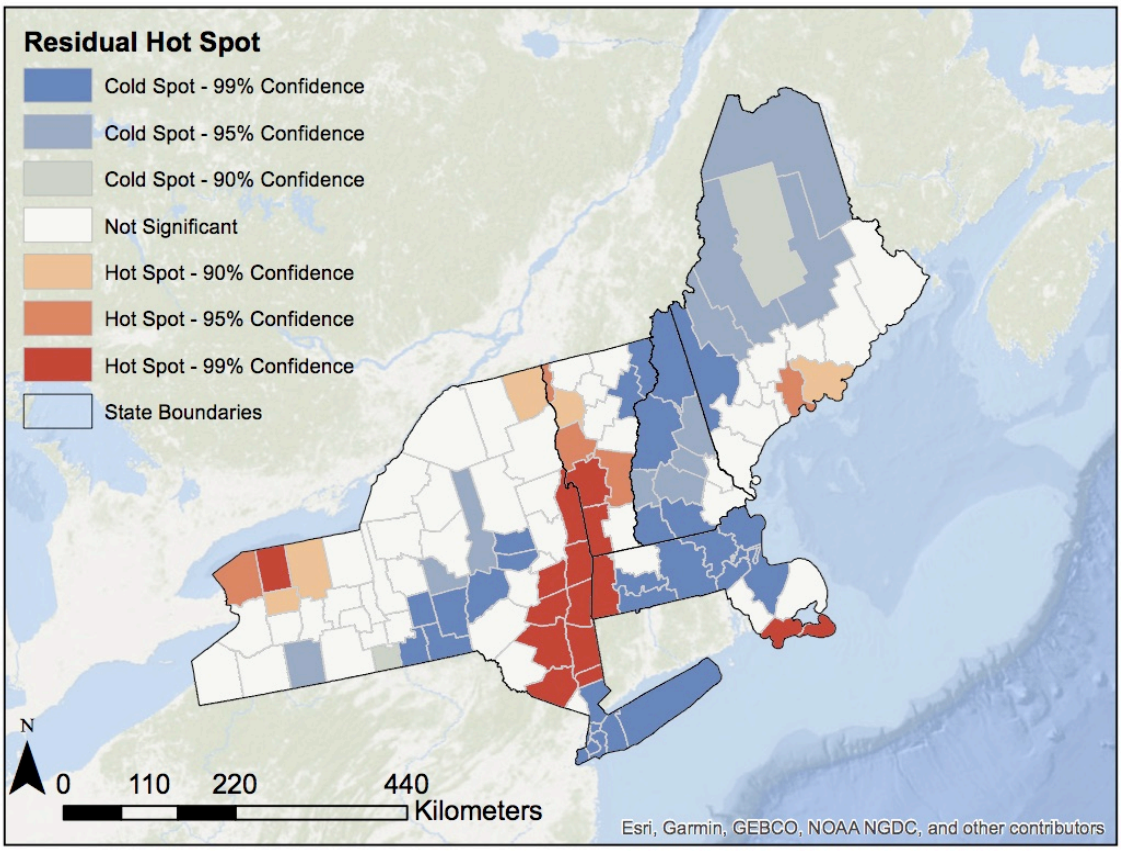

Figure 28. Residual Hot Spot Map for the study area

The dispersion of the residual hot spots showed the relationship with the Lyme disease rate. The areas marked as under-predicted (the hot spots) are very close to the locations marked as hot spots for the Lyme disease rate. Of the 25 counties marked as under-predictions, 14 of the counties were classified as hot spots in the Lyme disease rate hot spot map (Figure 15). The predicted values of the counties identified as hot spots in the Lyme disease hot spot maps was not expected to be accurate; however, the discrepancy was greater than expected. The over-predicted counties in the residual hot spot map do not correlate with the Lyme disease rates.

Overall, the final Lyme disease model showed Latitude with the greatest coefficient (-42.05), indicating lower latitudes correlated with a higher incidence rate. The Longitude of the counties was the second driver for the Lyme disease rate in the model, with a coefficient value of

12.4, indicating lower longitudes (or more eastern locations) correlate with a higher Lyme disease incidence. The mean temperature had a partial coefficient of -6.69 in the Lyme disease model, meaning that lower mean temperature correlates to higher incidence rates of Lyme disease. The variable year, which had a correlation coefficient of 3.56, was positively correlated with Lyme disease incidence and consistent with the increasing trend of Lyme disease incidence.

While the results of the residual analysis showed that the OLS 13 model had the best fit, the model only explained 14.4% of the variance in the Lyme disease rate. Overall, locations (latitude and longitude) were found to be the most influential variables for Lyme disease incidence. The mean temperature and the year were also influential, but were significantly less influential than location. Despite the lack of variance explained, the predicted Lyme disease map was created and the correlation between the variables and Lyme disease incidence were established for this study.

Chapter 5 Conclusions

The spatiotemporal analyses conducted in this study have provided vast insights into the correlation of Lyme disease and the selected environmental factors over space and time. This chapter provides a summary of the outcomes. Also included in this chapter are the major limitations in the study design with a discussion of future directions of the study.

The overall aim of the project, to understand the spatiotemporal relationships between Lyme disease and environmental factors, has been met. The rate of Lyme disease has previously been established as rising in the majority of the counties in the study area over the period of 2000 and 2015. This study found there were exceptions of to this trend. The exceptions were the declining trend of the Lyme disease in seven (7) counties. These counties, including Columbia County, Dutchess County, Nassau County, Orange County, Putnam County, Suffolk County, and Westchester County, were all located in the southeastern portion of New York. Within the seven counties, Columbia County was identified as a diminishing hotspot for the Lyme disease rate as well.

Three subregions of the study area contained the Lyme disease rate emerging hot spots. The largest area of the Lyme disease hot spot was centrally located at the juncture of New York, Vermont, and Massachusetts. It included six counties from New York, four counties from Vermont, and two counties from Massachusetts. The hot spots were a mix of consecutive hot spots, persistent hot spots, diminishing hot spots, and sporadic hot spots, indicating various changes in the temporal trend of Lyme disease in this subregion. The temporal neighbors are not consistent for each county throughout the study period.

The spread of the other two hot spot areas were confined to the eastern seaboard of the study area. Four counties in the southeastern portion of Maine were classified as consecutive hot

spots, indicating a statistically significant probability of Lyme disease in this area. Another four counties in the southernmost tip of Massachusetts also showed statistical significance in the probability of Lyme disease, but the temporal distributions vary.

The regression model of Lyme disease indicated precipitation and maximum temperature being the two significant environmental variables for Lyme disease incidence. When considering spatial (longitude and latitude) and temporal (year) variables, however, precipitation is no longer significant, and mean temperature became significant. The independent variables of our final Lyme disease model included year, longitude, latitude, and mean temperature. Spatial locations were overall the most critical drivers for the Lyme disease.

Because the R-squared of the Lyme disease model is fairly low (14.4%), this model is a suitable inferential model, as it indicated important variables that influence Lyme disease. Lower latitudes are correlated with higher incidence rates and more eastern locations (less negative longitude) are correlated with higher Lyme disease rates. In terms of climate, lower mean temperature attributes to the higher incidence rate of Lyme disease. Time (year) is positively correlated with Lyme disease incidence, which is consistent with the increasing temporal trend of Lyme disease incidence.

5.1 Limitations

There were several limitations in the project design and analysis. The main limitation was the use of aggregated data in the county level. Due to health privacy laws, the use of aggregated data is more accessible. However, the use of aggregated data decreases the power of the analysis. The spatial relationships of the Lyme disease incidence and the environmental factors might have been weakened due to this scale problem. The other data aggregation limitation is the aggregation of climate data. Climate data used for this study are spatially interpolated data from

weather stations. Using only one point per county as the climate variable might lessen the accuracy of the climate data. Unfortunately, the aggregation of county-level data did not allow a raster input for the entire interpolated area, which limited the actual distribution of climate data being included.

Another limitation is the overall data quality. The Lyme disease case counts as published by the CDC is unable to provide an accurate location or date of diagnosis. It also represented where the disease case was diagnosed but not where it was contracted. A case of Lyme disease can be contracted in one county, but diagnosed in another. A second data quality example is the lack of temporal variation of the forest cover data. It was necessary for the estimation of a land use variable, however the availability of merely one year of data hampered the estimation of the true temporal correlation. The final data quality issue is the missing data for few counties in the PRISM climate data. All of these limitations restricted the accuracy of the results, but were addressed as effectively as possible.

Another study limitation is the data accuracy from various sources. This is related to the data quality issue. The data used in the study were all secondary data that might contain unknown measurement errors and bias. This error problem might be reduced if the data collection was designed with the research objectives and done by the researcher. It was not possible for this study due to time constraints and lack of resources.

5.2 Future Directions and Implications

The study can be enhanced in multiple directions in the future. First, the study area can be expanded to include the states of Connecticut and Rhode Island. There are hotspots identified near the border of New York and Massachusetts, where Rhode Island and Connecticut are located. Including this spatial neighbor will help understanding the important spatiotemporal

trend of this area. The two states were excluded in this study for the concern of data processing time. By adding the total 14 counties of these two states, the number of county-year polygons to be analyzed would have increased to 224 instead of 116.

The next future direction will be to look at the areas that were notably different from other counties in the study. This can include any hot spot or Low-High Outlier areas. For example, by investigating the Low-High Outlier such as Delaware County, New York the reason behind the dynamics of Lyme disease rate can be further realized. The possible reasons could be either related to natural resources (e.g. fewer habitats available for ticks to survive), or related to human practices (e.g. better prevention methods).

Last but not least, a future direction can also be to investigate more variables related to tick habitat environment, particularly critical thresholds related to tick's survival. Determining the bounds of tick survivability would increase the ability to predict when an influx of Lyme disease may occur. When the environment is suitable for ticks, it would amplify the possible number of cases for that year and the year after. Studying the conditions of what a tick needs to survive, and what would kill a tick could inform possible tick epidemic in advance and prevent cases of Lyme disease from being contracted.

The Lyme disease models built in this study brought several benefits. It provides the guidance to public health officials in creating and distributing accurate preventative measures to the general public. The public health department can also benefit from this study to increase the knowledge and known qualities of Lyme disease. The spatiotemporal correlations between Lyme disease and the environmental factors identified in this study also can be used to educate the general public to be aware of the possible endemic areas and the environment of contracting Lyme disease.

References

Amrhein, Carl and Harold Reynolds. 1997. "Using the Getis Statistic to Explore Aggregation Effects in Metropolitan Toronto Census Data." *Canadian Geographer / Le Géographe Canadien* 41 (2): 137-149. Doi:10.1111/j.1541-0064.1997.tb01154.x. https://search.proquest.com/docview/228334468.

Barbour, Alan G. and Durland Fish. 1993. "The Biological and Social Phenomenon of Lyme Disease." *Science* 260: 1610-1616. Doi:10.1126/science.8503006. https://www.jstor.org/stable/3840183.

Brownstein, John S., Theodore R. Holford, and Durland Fish. 2003. "A Climate-Based Model Predicts the Spatial Distribution of the Lyme Disease Vector *Ixodes Scapularis* in the United States." *Environmental Health Perspectives* 111 (9): 1152-1157. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1241567/.

———. 2005. "Effect of Climate Change on Lyme Disease Risk in North America." *EcoHealth* 2 (1): 38-46. doi:10.1007/s10393-004-0139-x. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2582486/.

Centers for Disease Control and Prevention (CDC). 2017a. "Data and Statistics." Centers for Disease Control and Prevention, last modified November 13, 2017. https://www.cdc.gov/lyme/stats/index.html.

———. 2017b. "Lyme Disease Surveillance and Available Data." Centers for Disease Control and Prevention, last modified November 13, 2017. https://www.cdc.gov/lyme/stats/survfaq.html.

———. 2017c. "Signs and Symptoms of Untreated Lyme Disease." Centers for Disease Control and Prevention, last modified October 26, 2017. https://www.cdc.gov/lyme/signs_symptoms/index.html.

———. n.d. a. "Lyme Disease (*Borrelia Burgdorferi*) 1996 Case Definition." Centers for Disease Control and Prevention, accessed February 1, 2018, https://wwwn.cdc.gov/nndss/conditions/lyme-disease/case-definition/1996/.

———. n.d. b "Lyme Disease (*Borrelia Burgdorferi*) 2008 Case Definition." Centers for Disease Control and Prevention, accessed February 1, 2018, https://wwwn.cdc.gov/nndss/conditions/lyme-disease/case-definition/2008/**.**

Chen, Nengwang, Huancheng Li, and Lihong Wang. 2009. "A GIS-Based Approach for Mapping Direct use Value of Ecosystem Services at a County Scale: Management Implications." *Ecological Economics* 68 (11): 2768-2776. doi:10.1016/j.ecolecon.2008.12.001. https://www.sciencedirect.com/science/article/pii/S0921800908005296.

Ciesielski, Carol A., Lauri E. Markowitz, Rose Horsley, Allen W. Hightower, Harold Russell, and Claire V. Broome. 1989. "Lyme Disease Surveillance in the United States, 1983-1986." *Reviews of Infectious Diseases* 11 (Supplement 6): S1441. doi:10.1093/clinids/11.Supplement_6.S1435. http://www.jstor.org/stable/4455353.

Davis, Jeffrey P., Terry E. Amundson, Andrew Spielman, Alan G. Barbour, and Richard A. Kaslow. 1984. "Lyme Disease in Wisconsin: Epidemiologic, Clinical, Serologic, and Entomologic Findings." *The Yale Journal of Biology and Medicine* 57 (4): 685-696. https://www-ncbi-nlm-nih-gov.libproxy2.usc.edu/pmc/articles/PMC2590030/pdf/yjbm00100-0229.pdf.

Delmelle, Eric, Coline Dony, Irene Casas, Meijuan Jia, and Wenwu Tang. 2014. "Visualizing the Impact of Space-Time Uncertainties on Dengue Fever Patterns." *International Journal of Geographical Information Science* 28 (5): 1107-1127. doi:10.1080/13658816.2013.871285. https://search.proquest.com/docview/1521126311.

Diggle, P. J., A. G. Chetwynd, R. Häggkvist, and S. E. Morris. 1995. "Second-Order Analysis of Space-Time Clustering." *Statistical Methods in Medical Research* 4: 124-136.

Esri. 2018a. "How Emerging Hot Spot Analysis Works." Environmental Systems Research Institute. Accessed June 2018. http://pro.arcgis.com/en/pro-app/tool-reference/space-time-pattern-mining/learnmoreemerging.htm.

———. 2018b. "How Local Outlier Analysis Works." Environmental Systems Research Institute. Accessed June 2018. http://pro.arcgis.com/en/pro-app/tool-reference/space-time-pattern-mining/learnmorelocaloutlier.htm

———. 2018c. "Interpreting OLS Results." Environmental Systems Research Institute. Accessed June 2018. http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/interpreting-ols-results.htm

———. 2018d."Visualizing the Space Time Cube". Environmental Systems Research Institute. Accessed April 2018. http://pro.arcgis.com/en/pro-app/tool-reference/space-time-pattern-mining/visualizing-cube-data.htm

Fotheringham, A. 2010. "Geographically Weighted Regression." *Encyclopedia of Geography,* 1225-1232.

Fotheringham, Alexander Stewart, Chris Brunsdon, and Martin Charlton. 2002. Geographically Weighted Regression. Chichester: Wiley.

Frad, Wayne. 2017. "Capstone Proposal Penn State MGIS Program" Pennsylvania State.

Gilmore Jr, Robert D., M. Lamine Mbow, and Brian Stevenson. 2001. "Analysis of *Borrelia Burgdorferi* Gene Expression during Life Cycle Phases of the Tick Vector *Ixodes Scapularis*." *Microbes and Infection* 3: 799-808. doi:10.1016/S1286-4579(01)01435-6. https://www-sciencedirect-com.libproxy1.usc.edu/science/article/pii/S1286457901014356.

Glass, Gregory E., Brian E. Schwartz, John M. Morgan, Dale T. Johnson, Peter M. Noy, and Ebenezer Israel. 1995. "Environmental Risk Factors for Lyme Disease Identified with Geographic Information Systems." *American Journal of Public Health* 85 (7): 944-948.

Goodwin, Brett J., Richard S. Ostfeld, and Eric M. Schauber. 2001. "Spatiotemporal Variation in a Lyme Disease Host and Vector: Black-Legged Ticks on White-Footed Mice." *Vector Borne and Zoonotic Diseases* 1 (2): 129-138. doi:10.1089/153036601316977732. http://www.ncbi.nlm.nih.gov/pubmed/12653143.

Guerra, Marta, Edward Walker, Carl Jones, Susan Paskewitz, M. Roberto Cortinas, Ashley Stancil, Louisa Beck, Matthew Bobo, and Uriel Kitron. 2002. "Predicting the Risk of Lyme Disease: Habitat Suitability for *Ixodes Scapularis* in the North Central United States." *Emerging Infectious Diseases* 8 (3): 289-297. http://www.ncbi.nlm.nih.gov/pubmed/11927027.

Guo, Diansheng, Jin Chen, Alan M. MacEachren, and Ke Liao. 2006. "A Visualization System for Space-Time and Multivariate Patterns (VIS-STAMP)." *IEEE Transactions on Visualization and Computer Graphics* 12 (6): 1461-1474. doi:10.1109/TVCG.2006.84. http://ieeexplore.ieee.org/document/1703367.

Holt, James B., C.P. Lo, and Thomas W. Hodler. 2004. "Dasymetric Estimation of Population Density and Areal Interpolation of Census Data." *Cartography and Geographic Information Science* 31 (2): 103-121. doi:10.1559/1523040041649407. http://www.ingentaconnect.com/content/acsm/cagis/2004/00000031/00000002/art00004.

Homer, Collin, Jon Dewitz, Joyce Fry, Michael Coan, N. Hossain, C. Larson, Nate Herold, Alexa McKerrow, J.N. VanDriel, and James Wickham. 2007. "Completion of the 2001 National Land Cover Database for the Conterminous United States." *Photogrammetric Engineering and Remote Sensing*, Vol. 73, No. 4, pp 337-341.

Jelinski, Dennis E. and Jianguo Wu. 1996. "The Modifiable Areal Unit Problem and Implications for Landscape Ecology." *Landscape Ecology* 11 (3): 129-140. doi:10.1007/BF02447512.

Kitron, Uriel. 1998. "Landscape Ecology and Epidemiology of Vector-Borne Diseases: Tools for Spatial Analysis." *Journal of Medical Entomology* 35 (4): 435-445. doi:10.1093/jmedent/35.4.435. http://www.ingentaconnect.com/content/esa/jme/1998/00000035/00000004/art00015.

Kitron, Uriel and James J. Kazmierczak. 1997. "Spatial Analysis of the Distribution of Lyme Disease in Wisconsin." *American Journal of Epidemiology* 145 (6): 558-566. doi:10.1093/oxfordjournals.aje.a009145. http://www.ncbi.nlm.nih.gov/pubmed/9063347.

Kittayapong, Pattamaporn, Sutee Yoksan, Uruyakorn Chansang, Chitti Chansang, and Amaret Bhumiratana. 2008. "Suppression of Dengue Transmission by Application of Integrated Vector Control Strategies at Sero-Positive GIS-Based Foci." *American Society of Tropical Medicine and Hygiene* 78 (1): 70-76. http://www.ajtmh.org/cgi/content/abstract/78/1/70.

Kjellin, Andreas, Lars Winkler Pettersson, Stefan Seipel, and Mats Lind. 2010. "Different Levels of 3D: An Evaluation of Visualized Discrete Spatiotemporal Data in Space-Time Cubes." *Information Visualization* 9 (2): 152-164. doi:10.1057/ivs.2009.8. http://dx.doi.org/10.1057/ivs.2009.8

LoGiudice, Kathleen, Shannon T. K. Duerr, Michael J. Newhouse, Kenneth A. S. Schmidt, Mary E. Killilea, and Richard S. Ostfeld. 2008. "Impact of Host Community Composition on Lyme Disease Risk." *Ecology* 89 (10): 2841-2849. doi:10.1890/07-1047.1. http://www.jstor.org/stable/27650829.

Longley, Paul A. 2012. "Geodemographics and the Practices of Geographic Information Science." *International Journal of Geographical Information Science* 26 (12): 2227-2237. doi:10.1080/13658816.2012.719623. http://www.tandfonline.com/doi/abs/10.1080/13658816.2012.719623.

Multi-Resolution Land Characteristics Consortium. 2017. *NLCD Evaluation, Visualization, and Analysis (EVA) Tool.* U.S. Department of Interior, Unites States Geological Service.

Nakaya, Tomoki. 2013. "Analytical Data Transformations in Space–Time Region: Three Stories of Space–Time Cube." *Annals of the Association of American Geographers* 103 (5): 1100-1106. doi:10.1080/00045608.2013.792184. https://search.proquest.com/docview/1427374600.

Nicholson, Matthew C. and Thomas N. Mather. 2014. "Methods for Evaluating Lyme Disease Risks using Geographic Information Systems and Geospatial Analysis." *Journal of Medical Entomology* 33 (5): 711-720.

Needham, Glenn R. and Pete D. Teel. 1991. "Off-Host Physiological Ecology of Ixodid Ticks." *Annual Review of Entomology* 36 (1): 659-681.

O'Sullivan, David and David Unwin. 2010. *Geographic Information Analysis, 2nd Edition*. Wiley.

Ogden, Nicholas H., Catherine Bouchard, Klaus Kurtenbach, Gabriele Margos, L.R. Lindsay, Louise Trudel, Soulyvane Nguon, and François Milord. 2010. "Active and Passive Surveillance and Phylogenetic Analysis of *Borrelia Burgdorferi* Elucidate the Process of Lyme Disease Risk Emergence in Canada." *Environmental Health Perspectives* 118 (7): 909-914. doi:10.1289/ehp.0901766. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2920908/.

Oliver, James H. 1996. "Lyme Borreliosis in the Southern United States: A Review." *The Journal of Parasitology* 82 (6): 926-935. doi:10.2307/3284201. http://www.jstor.org.libproxy1.usc.edu/stable/3284201.

Orloski, Kathleen A., Edward B. Hayes, Grant L. Campbell, and David T. Dennis. 2000. "Surveillance for Lyme Disease--United States, 1992-1998." *MMWR. CDC Surveillance Summaries : Morbidity and Mortality Weekly Report. CDC Surveillance Summaries / Centers for Disease Control* 49 (3): 1-11. http://www.ncbi.nlm.nih.gov/pubmed/10817483.

Palmer, Michael W. and Peter S. White. 1994. "Scale Dependence and the Species-Area Relationship." *The American Naturalist* 144 (5): 717-740. doi:10.1086/285704. http://www.jstor.org/stable/2463009.

PRISM Climate Group. "PRISM Climate Data." Oregon State University, Last modified 2017, Accessed February 2018, http://prism.oregonstate.edu/.

Reed, R. A., Robert K. Peet, Michael W. Palmer, and P. S. White. 1993. "Scale Dependence of Vegetation-Environment Correlations: A Case Study of a North Carolina Piedmont Woodland." *Journal of Vegetation Science* 4 (3): 329-340. doi:10.2307/3235591. http://www.jstor.org/stable/3235591.

Schuurman, Nadine, Nathaniel Bell, James R. Dunn, and Lisa Oliver. 2007. "Deprivation Indices, Population Health and Geography: An Evaluation of the Spatial Effectiveness of Indices at Multiple Scales." *Journal of Urban Health* 84 (4): 591-603. doi://dx.doi.org/10.1007/s11524-007-9193-3.

Spielman, Andrew. 1994. "The Emergence of Lyme Disease and Human Babesiosis in a Changing Environment." *Annals of the New York Academy of Sciences* 740 (1): 146-156. doi:10.1111/j.1749-6632.1994.tb19865.x.

Steere, Allen C., Stephen E. Malawista, David R. Snydman, Robert E. Shope, Warren A. Andiman, Martin R. Ross, and Francis M. Steele. 1977. "Lyme Disease: An Epidemic of Oligoarticular Arthritis in Children and Adults in Three Connecticut Communities." *Arthritis & Rheumatism* 20 (1): 7-17. doi:10.1002/art.1780200102. http://dx.doi.org/10.1002/art.1780200102.

Steere, Allen C., Thomas F. Broderick, and Stephen E. Malawista. "Erythema Chronicum Migrans and Lyme Arthritis: Epidemiologic Evidence for a Tick Vector." *American Journal of Epidemiology* 108, no. 4 (1978): 312-321.

Szwarcwald, Célia L., Francisco Inácio Bastos, Christovam Barcellos, Maria de Fátima Pina, and Maria Angela Pires Esteves. 2000. "Health Conditions and Residential Concentration of Poverty: A Study in Rio De Janeiro, Brazil." *Journal of Epidemiology and Community Health* 54 (7): 530-536. doi:10.1136/jech.54.7.530. http://www.jstor.org/stable/25569233.

Thomson, Madeleine C. and Stephen J. Connor. 2000. "Environmental Information Systems for the Control of Arthropod Vectors of Disease." Medical and Veterinary Entomology 14 (3): 227-244. https://pdfs.semanticscholar.org/0025/73811ff089cd4796f55ab850ce99c9bd8ee9.pdf.

Tominski, Christian, Petra Schulze-Wollgast, and Heidrun Schumann. 2005. "3D Information Visualization for Time Dependent Data on Maps." IEEE.

United States Census Bureau. 2016. “TIGER Geodatabase”, Accessed February 2018, https://www.census.gov/geo/maps-data/data/tiger-geodatabases.html.

Waller, Lance A., Brett J. Goodwin, Mark L. Wilson, Richard S. Ostfeld, Stacie L. Marshall and Edward B. Hayes. 2007. "Spatio-Temporal Patterns in County-Level Incidence and Reporting of Lyme Disease in the Northeastern United States, 1990–2000." *Environmental and Ecological Statistics* 14 (1): 83-100. doi:10.1007/s10651-006-0002z. https://search.proquest.com/docview/756947797.

Zundel, Julie. "Ticks: Life Cycle & Reproduction" Study.com, Accessed July 17, 2018, https://study.com/academy/lesson/ticks-life-cycle-reproduction.html.

Appendix A Lyme Disease Rate Hot Spot Classifications

County	State	Classification	Hot Spot
Hancock County	Maine	Consecutive Hot Spot	Eastern (2)
Knox County	Maine	Consecutive Hot Spot	Eastern (2)
Lincoln County	Maine	Consecutive Hot Spot	Eastern (2)
Waldo County	Maine	Consecutive Hot Spot	Eastern (2)
Barnstable County	Massachusetts	Sporadic Hot Spot	Southern (3)
Berkshire County	Massachusetts	Diminishing Hot Spot	Central (1)
Bristol County	Massachusetts	Consecutive Hot Spot	Southern (3)
Dukes County	Massachusetts	Intensifying Hot Spot	Southern (3)
Franklin County	Massachusetts	Sporadic Hot Spot	Central (1)
Nantucket County	Massachusetts	Intensifying Hot Spot	Southern (3)
Albany County	New York	Persistent Hot Spot	Central (1)
Columbia County	New York	Diminishing Hot Spot	Central (1)
Greene County	New York	Persistent Hot Spot	Central (1)
Rensselaer County	New York	Persistent Hot Spot	Central (1)
Ulster County	New York	Diminishing Hot Spot	Central (1)
Washington County	New York	Consecutive Hot Spot	Central (1)
Bennington County	Vermont	Consecutive Hot Spot	Central (1)
Rutland County	Vermont	Consecutive Hot Spot	Central (1)
Windham County	Vermont	Consecutive Hot Spot	Central (1)
Windsor County	Vermont	Consecutive Hot Spot	Central (1)

Appendix B Ordinary Least Squares Models 1 to 13

	Variables	ß Sign	ß Sig	VIF	Jarque Bera Sig	A R^2	AICC	Residuals	Koenker Sig
OLS 1	Ppt, Min T, Mean T, Max T	No	No	No	Yes	0.0290	22546.035	Yes	No
OLS 2	Ppt, Mean T, Max T	No	No	No	Yes	0.0291	22544.890	Yes	No
OLS 3	Ppt, Max T	Yes	Yes	Yes	Yes	0.0288	22544.586	Yes	No
OLS 4	Year, Longitude, Latitude, Ppt, Min T, Mean T, Max T, Forest Cover	No	No	No	Yes	0.146	22312.848	Yes	No
OLS 5	Year, Longitude, Latitude, Ppt, Mean T, Max T, Forest Cover	No	No	No	Yes	0.146	22310.362	Yes	No
OLS 6	Year, Longitude, Latitude, Ppt, Max T, Forest Cover	No	No	Yes	Yes	0.137	22329.543	Yes	No
OLS 7	Year, Longitude, Latitude, Ppt, Mean T, Forest Cover	No	No	Yes	Yes	0.144	22314.274	Yes	No
OLS 8	Year, Longitude, Latitude, Min T, Mean T, Max T, Forest Cover	No	No	No	Yes	0.146	22310.042	Yes	No
OLS 9	Year, Longitude, Latitude, Mean T, Max T, Forest Cover	Yes	No	Yes	Yes	0.147	22308.603	Yes	No

OLS 10	Year, Longitude, Latitude, Mean T, Forest Cover	Yes	No	Yes	Yes	0.144	22312.281	Yes	No
OLS 11	Year, Longitude, Latitude, Max T, Forest Cover	Yes	No	Yes	Yes	0.137	22327.758	Yes	No
OLS 12	Year, Longitude, Latitude, Max T,	Yes	Yes	Yes	Yes	0.137	22328.280	Yes	No
OLS 13	Year, Longitude, Latitude, Mean T	Yes	Yes	Yes	Yes	0.144	22312.169	Yes	No

ß - partial coefficient
Sig - Significant
Ppt - Precipitation