2017

Spatial and Spatiotemporal Modeling of Epidemiological Data

Laxman Karki
South Dakota State University

SPATIAL AND SPATIOTEMPORAL MODELING OF EPIDEMIOLOGICAL DATA

BY

LAXMAN KARKI

A dissertation submitted in partial fulfillment of the requirements for the

Doctor of Philosophy

Major in Computational Science and Statistics

South Dakota State University

2017

SPATIAL AND SPATIOTEMPORAL MODELING OF EPIDEMIOLOGICAL DATA

This dissertation is approved as a creditable and independent investigation by a candidate for the Doctor of Philosophy in Computational Science and Statistics degree and is acceptable for meeting the dissertation requirements for this degree. Acceptance of this does not imply that the conclusions reached by the candidate are necessarily the conclusions of the major department.

Gary D. Hatfield, Ph.D. Date

Dissertation Advisor

Kurt Cogswell, Ph.D. Date

Head, Department of Mathematics and Statistics

Dean, Graduate School Date

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor Dr. Gary D. Hatfield for his invaluable support, inspiration, suggestions, and guidance throughout this research and my graduate studies.

I would like to thank Dr. Thomas Roe, and Dr. Thomas Brandenburger for serving as the member of my Ph.D. committee, continuous support, and constructive advice during the preparation of this dissertation. I would also like to thank Dr. Billy Fuller for his time and effort as a Graduate Faculty Representative of my dissertation committee.

I would like to thank Dr. Kurt Cogswell for supporting my graduate studies with a graduate teaching assistantship, and Dr. Donna Flint for her help in my journey of teaching undergraduate courses at SDSU. I would also like to extend my appreciation to the faculty and staff in Department of Mathematics and Statistics at SDSU.

My sincere gratitude goes to my parents, siblings, and all other family members. Their continuous love and support throughout the years are highly remarkable.

I would like to acknowledge my little boy Sujay whose smile kept me motivated. It's hard to put into words how much I love you.

Finally, but most importantly, I would like to thank my wife Sajag Adhikari. Without her unending support, encouragement and love, I could have never completed the Ph.D. program.

TABLE OF CONTENTS

ABBREVIATIONS

AIC: Akaike Information Criterion

AICc: Akaike Information Criterion Corrected

BIC: Bayesian Information Criterion

BYM: Besag-York-Mollie

BRFSS: Behavioral Risk Factor Surveillance System

CDC: Centers for Disease Prevention and Control

CV: Cross Validation

DIC: Deviance Information Criterion

GMRF: Gaussian Markov Random Field

GPS: Global Positioning System

GWR: Geographically Weighted Regression

iCAR: Intrinsic Conditional Autoregressive

INLA: Integrated Nested Laplace Approximations

IQR: Inter Quartile Range

LISA: Local Indicators of Spatial Association

MCMC: Markov Chain Monte Carlo

MGWR: Mixed Geographically Weighted Regression

MLE: Maximum Likelihood Estimate

OLS: Ordinary Least Squares

ZINB: Zero-Inflated Negative Binomial

ZIP: Zero-Inflated Poisson

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

SPATIAL AND SPATIOTEMPORAL MODELING OF EPIDEMIOLOGICAL DATA

LAXMAN KARKI

2017

This dissertation focuses on modeling approach for spatial and spatiotemporal data with epidemiological applications. Chapter one gives the general overview of spatial and spatiotemporal data and challenges in the statistical analysis of spatial and spatiotemporal data, and motivation and objectives of the study.

Chapter two describes the regression models commonly used in spatial data analysis. Various types of regression methods such as OLS, GWR and MGWR were used to study the association between diabetes prevalence and socioeconomic and lifestyle factors on county level data of Midwestern United States. A new analysis workflow is purposed for regression analysis of spatial data.

Chapter three describes recently developed INLA as an alternative of traditionally used MCMC in Bayesian hierarchical models. INLA method was used to identify the best regression model for the spatiotemporal regression analysis of Lyme disease count data with climatic covariates in county-level data in Minnesota.

Chapter four gives the contribution of this dissertation and discusses the direction for the future research.

Chapter 1

1 General Introduction

1.1 Background

The availability of spatial and spatiotemporal data has increased substantially in the last few decades due to advancement in computational tools, which enables us to collect real-time data coming from GPS, satellite etc. (Cressie, 2015; Plant, 2012; Ripley, 2005). Researchers nowadays in a wide variety of fields including epidemiology, forestry, and sociology to hydrology, have to deal with spatial and spatiotemporal data. Spatial data constitutes information about both an attribute of interest as well as its location. The location may include a set of coordinates such as longitude and latitudes or small areas such as census tracts, counties etc.

An example we can consider is an epidemiologist to evaluate the incidence of particular diseases such as Lyme disease in some state or geographical regions. The data is usually available in counts of people infected with the disease in small areas, for example, county-level count data of Lyme disease over the years. Researchers can answer many questions for example: is there a potential geographical pattern of the disease for areas close to each other that have similar incidence? Is there some temporal pattern of the disease?

According to Blangiardo et al. (2013), spatial data are defined as realized values of stochastic process indexed by space as:

$$\mathbf{Y(s)} \equiv \{y(\mathbf{s}), \mathbf{s} \in \mathbf{D}\}$$

Where **D** is a fixed subset of $\mathbf{R^d}$ (Here we consider d=2). A collection of observations **y** = (y ($\mathbf{s_1}$), y ($\mathbf{s_2}$),…, y ($\mathbf{s_n}$)) represents the actual data and ($\mathbf{s_1, s_{2,\ldots,} s_n}$) represents the spatial units of measurements. If **D** is continuous surface then the problem can be specified as spatially continuous random process, and if **D** is a countable collection of d-dimensional spatial units then the problem is defined as discrete random process (Blangiardo, Cameletti, Baio, & Rue, 2013; Gelfand, Diggle, Guttorp, & Fuentes, 2010).

We can specify spatial data into three different categories as:

1.1.1 Areal or Lattice data: Lattice refers to a situation where y(**s**) is the aggregation of values over areal units such as zip codes, counties(s) with well-defined boundaries in **D**. For example, we can aggregate all the cases of Lyme disease per counties in Minnesota. The difference between areal and lattice data is former is irregular in shape and the boundaries are defined based on administrative boundaries such as postal code, census tract, counties etc. whereas later is regular in shape, for example, we can collect the number certain plant species present in regularly shaped areal quadrats.

1.1.2 Point-referenced (Geostatistical) data: This category consists of data measured at specific location y(**s**) where the spatial domain **s** varies continuously over the spatial domain **D**. The location **s** is commonly represented by two- dimensional vector longitude and latitude. The actual data are represented by observations **y** = (y ($\mathbf{s_1}$), y ($\mathbf{s_2}$), …, y ($\mathbf{s_n}$)) at locations ($\mathbf{s_1, s_2,\ldots,s_n}$). For example, we can get the measurement of temperature and precipitation data from weather stations located at different locations. A common goal of point-referenced (Geostatistical) data is to interpolate y where the data measurements are not available.

1.1.3 Point pattern data: The point pattern data y(s) is the collection of information about whether the event of interest occurred or not at random locations. The spatial domain **D** represents the set of points where the event occurred. For example, we might be interested in locations of nests of a bird species in a forest or addresses of persons with a certain disease. In these examples, the location **S** in $\mathbf{R^d}$ is random and the measurements y(**s**) are taken as binary value 0, 1 based on whether the event has occurred or not. The main question of interest with point pattern data is whether the event of interest is random or clustered in the spatial domain **D**.

1.2 Spatiotemporal data:

Spatiotemporal data is a simple extension of spatial data with adding time dimension. Spatiotemporal data are defined as:

$$\mathbf{Y}(s,t) \equiv \left\{\mathbf{y}(s,t), (s,t) \in \mathbf{D} \in \mathbf{R}^2 \times \mathbf{R}\right\}$$

Where data is observed in n spatial areas or locations and at T time points. Tobler's first law of geography states that "everything is related to everything else but near things are more related than distant things" (Tobler, 1970). Spatial data are usually correlated either positively or negatively with proximal locations. Positive spatial autocorrelation arises if similar values cluster together in a map; similarly, negative spatial autocorrelation arises when dissimilar values cluster together in map. There is the presence of autocorrelation due to spatial dependency. The use of standard statistical techniques, which assumes independence of observations, are not appropriate for spatial data due to spatial

autocorrelation. Figure 1.1 displays an example of positive and negative spatial autocorrelation.

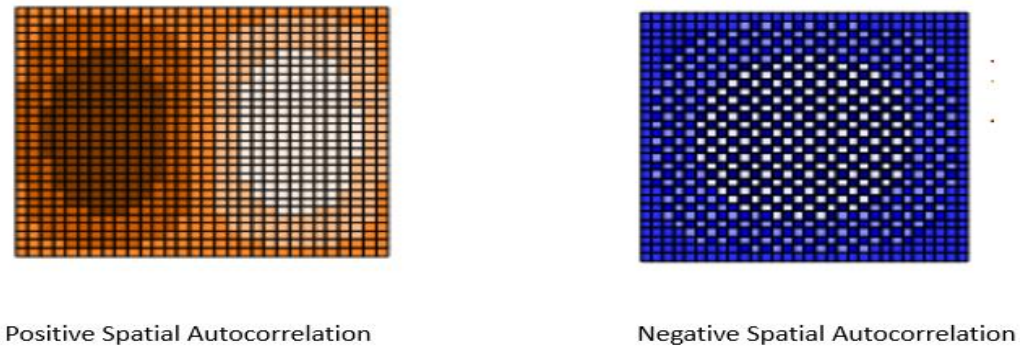


Figure 1. 1 An example of positive and negative spatial autocorrelation (Source: Dr. Ronald Briggs, with modification).

The relationship between variables might be different at different points in space. The use statistical models in spatial data without considering the possibilities of variation of effects with geographical locations commits Simpson's paradox which is the 'reversal of results when groups of data are analyzed separately and then combined (Fotheringham, Brunsdon, & Charlton, 2003). The effect of Simpson's paradox in spatial data analysis is better displayed by figure 1.2:

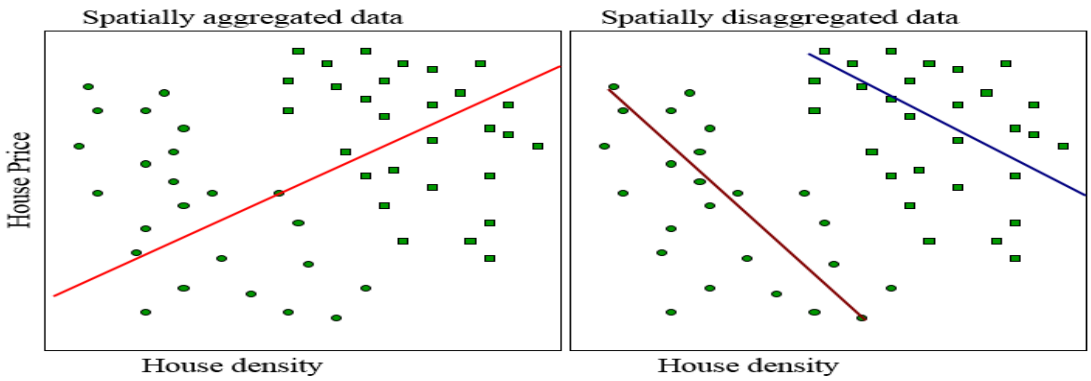


Figure 1. 2 Example of Simpson's paradox.

Source: www.pages.csam.montclair.edu/~yu/GISDay_GWR.ppt

1.3 Measurement of Spatial Autocorrelation

1.3.1 Global Indices of Spatial Autocorrelation

Spatial autocorrelation is the correlation of the same measurement taken at different areal units. Global indices of spatial autocorrelation are to summarization of degree to which similar observation tend to occur near each other. It gives the summary over the entire region rather than a test to detect local spatial clusters. It calculates the similarity of values at location i and j and then weights the similarity by the proximity of locations i and j. High similarities with high weight are the indication of similar values that are closer together and low similarities with high weight indicate dissimilar values that are close together. The value of global spatial autocorrelation help to summarize the similarity of nearby areal units. The similarities of values A_i and A_j are weighted by the proximities of i and j. The weight w_{ij} defines proximity. The weighted average of similarities between areal units represents the extent of similarities. Global indices of spatial autocorrelation are built on this basic form:

$$\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}$$

1.3.2 Moran's I:

Moran's I is the basic extension of global indices of local autocorrelation. The similarity between areal units i and j is defined as the product of the respective difference between y_i and y_j with the overall mean divided by sample variance as:

$$Moran's\ I = \frac{n}{S_0} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(yi - \bar{y})^2}$$

Where S_0 is standard deviation.

The value of Moran's I varies in the interval [-1, 1]. We can interpret the value as similar to correlation coefficients. When the neighboring regions tend to have similar values then the value of Moran's I will be positive and when the neighboring regions have dissimilar values then Moran's I will be negative.

1.3.3 Calculation of weight matrices

Most of the spatial models are based on whether one region is the spatial neighbor of another region. Weight matrix is a square symmetric n x n matrix with (i,j) element is equal to 1 if region i and j are neighbors of one another, and zero otherwise. The diagonal elements of the spatial weight matrix are zeros. The most common ways to construct such a matrix are as follows:

(1) Rook case contiguity: Two regions are spatial neighbors if they share a common border (on any side). In this case, two regions are considered as neighbors if that border is longer than predefined small "snap distance". Figure 1.3 gives an example of Rook's case.

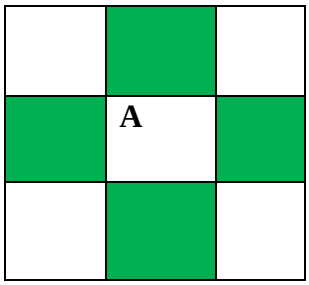


Figure 1. 3 Rook's case contiguity.

(2) Bishop case contiguity: two spatial regions meet at a point. This is similar to two elements of a graph meeting at a vertex. Bishop contiguity case arises when two regions share a common border and that is shorter than "snap distance". Figure 1.4 gives an example Bishop's case contiguity.

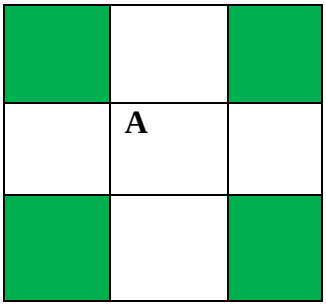


Figure 1. 4 Bishop's case contiguity.

Queen's case contiguity: It is combination of Rook's case and Bishop's case. Figure 1.5 gives an example of Queen's case contiguity.

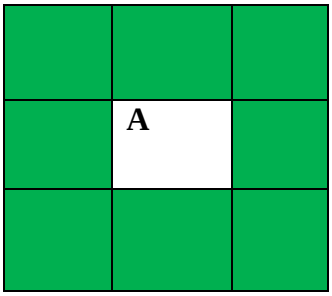


Figure 1. 5 Queen's case contiguity.

1.3.4 Spatial weight matrix (with Rook's case):

A	B
C	D

As described earlier, the elements of the weights matrix are 1 if the neighbors share the border and zero otherwise in Rook's case. Suppose we have four spatial lattice A, B, C and D as shown in the figure above. The weight matrix is calculated as:

$$W_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

Usually, spatial weights matrix is row standardized by dividing each element of the weight matrix by the corresponding row sum as:

$$W_{ij} = \begin{pmatrix} 0 & 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0 & 0.5 \\ 0.5 & 0 & 0 & 0.5 \\ 0 & 0.5 & 0.5 & 0 \end{pmatrix}$$

1.4 Global Measures of Spatial Autocorrelation

1.4.1 Geary's C

Geary's C statistic (Geary 1954) is based on the deviations in responses of each observation with one another. We can calculate Geary's C value as:

$$\text{Geary's } C = \frac{n-1}{2S_0} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - y_j)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

The value of Geary's C lies in the range [0, 2]. The value of 1 means there is no spatial autocorrelation. Values less than 1 indicates there is increasing positive spatial autocorrelation and values higher than 1 illustrate increasing negative spatial autocorrelation. There is an inverse relationship between Geary's C and Moran's I but they are not identical.

Moran's I is more global measurement and more sensitive to extreme values of y while Geary's C is more sensitive to local spatial autocorrelation. They both are asymptotically normally distributed as n increases (A. D. Cliff & Ord, 1981). In general, Moran's I and Geary's C results in the same conclusion regarding spatial autocorrelation, however, Moran's I is more powerful than Geary's C (A. Cliff & Ord, 1975; A. D. Cliff & Ord, 1981).

1.5 A local measure of spatial autocorrelation:

One might expect that sub-regions of a greater whole could have different local autocorrelation than that characterized by the single statistic that describes the entire region (Plant, 2012). The strength of global Moran's I is its simplicity in calculations and interpretation. The major limitation is that it takes the average local variations in the strength of global spatial autocorrelation. To overcome this problem, statisticians have developed local indices of spatial autocorrelation. The statistical methods to examine the

local level of spatial autocorrelation is very helpful in order to identify areas where values of the variable are extreme and geographically homogeneous (Anselin, 1995).

Anselin (1995) developed a standard tool, local indicator of spatial autocorrelation (LISA) to examine the local autocorrelation. It is the local equivalent of Moran's I. The sum of all the indices is proportional to the global value of Moran's statistics.

1.5.1 Anselin Local Moran's I:

Anselin local Moran's I statistic of spatial autocorrelation is calculated as

$$I_i = \frac{y_i - \bar{y}}{S_i^2} \sum_{j=1, j\neq i}^{n} w_{ij}(y_j - \bar{y})$$

$\sum_i I_i = I$, and I= Global Moran's I value.

Where y_i and $\bar{y}$ are the attribute of the feature i and the mean of the corresponding attribute respectively. W_{ij} is weight matrix.

$$S_i^2 = \frac{\sum_{j=1, j\neq i}^{n} (y_j - \bar{y})^2}{n-1}$$

And $$Z_{I_i} = \frac{I_i - E[I_i]}{\sqrt{V[I_i]}}$$

Where

$$E[I_i] = -\frac{\sum_{j=1, j\neq i}^{n} w_{ij}}{n-1}$$

$$V[I_i] = E[I_i^2] - E[I_i]^2$$

A significantly low p-value of the test statistic gives statistically significant cluster. Permutations are used to determine the significant cluster. If there is spatially significant pattern in data, local Moran's I values generated from permutation display less clustering than the local Moran's I value from the original data.

LISA values for each location allow to compute its similarity with its neighbors and to test its significance. The test can result in five different scenarios as:

(1) Hot spots: Locations with high values with similar locations. High-high.
(2) Cold spots: Locations with low values with similar locations. Low-low.
(3) Potential spatial outliers: Locations with high values with low value locations. High-low.
(4) Potential spatial outliers: Locations with low values with high value locations. Low-high.
(5) Locations with no significant local autocorrelations.

There is also presence of spatial heterogeneity, which is due to locational effect. Overall parameter estimates for entire region may not describe the process at any given location.

Spatial statistical methods have been used in wide variety of research areas: epidemiology, forestry, ecology, urban planning, and so many other research areas. The focus of this research was the application of spatial statistical methods in epidemiological data. The research objectives of my studies are as follows:

1.6 Objectives

(1) To determine the spatial prevalence of diabetes and how the distribution is associated with the geography of socio demographic and life style covariates in the Midwestern United States County-level diabetes data.

(2) To determine the spatial pattern of human cases of Lyme disease in Minnesota.

(3) To estimate the relationship between Lyme disease count and environmental risk factors using spatiotemporal methods in Minnesota.

1.6.1 Motivation and significance of objective 1

Most of the studies considered GWR model for regression analysis of spatial data. Even though GWR model ignores the possibility of having global and local effects of covariates. Fotheringham, Brunsdon, & Charlton (2003) purposed MGWR model to accommodate local and global regression coefficients in a single model. There are a plethora of research publications on spatial epidemiology that fitted GWR model to find the relationship between variables. To the best of my knowledge, none of the studies have considered MGWR model in spatial epidemiology.

Diabetes is a serious health threat with an alarming increase in prevalence rate among general population globally (Barker, Kirtland, Gregg, Geiss, & Thompson, 2011; Dijkstra et al., 2013; Kauhl et al., 2015; Wild, Roglic, Green, Sicree, & King, 2004). Diabetes and its complications are one of the serious health concern in the United States (Hipp & Chalise, 2015; Shaw, Sicree, & Zimmet, 2010). Diabetes prevalence rate for persons of age group 65 years or older is more than 10 times higher than people of age group 45 or younger (Engelgau, Geiss, Saaddine, & et al., 2004). According to CDC 2012 report, there were

29.1 million Americans, 9.3 percent of the total population had diabetes with estimated total health care costs of $ 245 billion in 2012. Type 2 diabetes which accounts for more than 90 percent of total diabetes affects people of all sex, age, race and ethnic groups however the rate is higher in American Indians, African Americans and people with socioeconomic disadvantages (Haire-Joshu, 2015).

This study will fit MGWR model as proposed by (Fotheringham et al., 2003) with county-level diabetes prevalence data for the Midwestern United States. This study will help to understand the relationship between sociodemographic covariates and diabetes prevalence in the Midwestern United States. The findings of this study will help public health programs to better target populations at risk of diabetes.

1.6.2 Motivation and Significance of study of objective 2

Lyme disease is one of the most frequent vector born disease in the United States (Killilea, Swei, Lane, Briggs, & Ostfeld, 2008; Orloski, Hayes, Campbell, & Dennis, 2000). It is expanding geographically and in its severity of impact (Hanrahan et al., 1984; Schauber & Ostfeld, 2002; Allen C Steere, Coburn, & Glickstein, 2004; Allen C Steere, Taylor, Wilson, Levine, & Spielman, 1986). The disease is transmitted by "black-legged" tick, *Ixodes scapularis* or *Ixodes pacificus* (A. C. Steere, Hardin, & Malawista, 1978). The early symptoms of Lyme disease are fever, skin rash, headache, and fatigue. If the disease is not treated in the initial stage of infection, more severe complications such as arthritis in major joints, intense pain, numbness or tingling in the hands or feet, and memory problem (Li et al., 2014). A 1998 study estimated the financial burden of Lyme disease in the United States were about $2.8 billion over a 5- year period (Maes, Lecomte, & Ray, 1998). According to CDC 2014 report, 96 percentage of confirmed Lyme disease cases is reported

to 14 states: Connecticut, Delaware, Maine, Maryland, Massachusetts, Minnesota, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont, Virginia, and Wisconsin. Lyme disease is distributed unevenly in every spatial scale and the vast majority of confirmed cases are in the northeastern and the upper Midwestern part of the United States and coastal central/northern California (Orloski et al., 2000).

Minnesota is one of the states in the United States with the high incidence of Lyme disease cases. To the best of my knowledge, there is no systematic study of Lyme disease distribution focusing on the state of Minnesota. This study aims to focus on spatial cluster analysis to detect Lyme disease clusters in Minnesota.

The emergence of infectious disease over past several decades highlights the needs of better understandings to handle the challenge of being epidemics or spread the disease to new locations. The understanding of spatial and temporal pattern of this disease will help to prepare and allocate the public health resources for disease prevention and control.

1.6.3 Motivation and significance of objective 3

The availability of data with spatial and temporal component has increased dramatically in the last few years. Some of the epidemiological data have the outcome and the risk factor characterized by the spatial and temporal structure which need to be considered for the inferential process (Blangiardo et al., 2013). Spatial data often recorded as count in certain spatial units. For example, in epidemiology, the number of infected people of certain disease per spatial units (states, counties) are available. Such data with information over time is also very common.

Climatic factor plays important role in tick borne diseases (Ogden et al., 2008; Raghavan, Goodin, Neises, Anderson, & Ganta, 2016; Subak, 2003). Climate and weather condition plays a key role in the incidence of Lyme disease because demography and distribution of *I. scapularis* are sensitive to variation in temperature and precipitation (Burtis et al., 2016; Eisen, Eisen, & Beard, 2016). Tick requires relatively humid microclimate. A recent study in northern Illinois found a significant relationship between cumulative rainfall and tick infection rates (Jones & Kitron, 2000).

The challenge of modeling spatiotemporal count data is the presence of many zeros and spatiotemporal correlation. A zero count is due to complete absence of the persons with Lyme disease in a given year or it might be the result of incomplete survey or imperfect detection. The zero-inflated Poisson model better in modeling such data (Agarwal, Gelfand, & Citron-Pousty, 2002; Wang, Chen, Kuo, & Dey, 2015).

Lyme disease count data follows a non-Gaussian distribution. Statistical modeling of count data is challenging due to counties with zero counts and there is complicated spatiotemporal dependence. Bayesian approach is effective in spatiotemporal data analysis. Bayesian hierarchical models can be implemented to estimate the parameters of spatiotemporal data (Lawson, 2013; Musenge, Chirwa, Kahn, & Vounatsou, 2013). Application of Bayesian hierarchical models in spatiotemporal analysis is challenging. Implementation of Bayesian hierarchical model relies on computationally expensive MCMC simulation techniques. A recently developed INLA has been an effective alternative of computationally expensive MCMC.

This study will find the relationship between Lyme disease count and climatic risk factors by using Bayesian hierarchical models in INLA. The findings of this study will help to understand the effects of climatic risk factors with Lyme diseases cases over time.

References

Agarwal, D. K., Gelfand, A. E., & Citron-Pousty, S. (2002). Zero-inflated models with application to spatial count data. *Environmental and Ecological Statistics, 9*(4), 341-355. doi:10.1023/a:1020910605990

Anselin, L. (1995). Local indicators of spatial association—LISA. *Geographical Analysis, 27*(2), 93-115.

Barker, L. E., Kirtland, K. A., Gregg, E. W., Geiss, L. S., & Thompson, T. J. (2011). Geographic distribution of diagnosed diabetes in the US: a diabetes belt. *American journal of preventive medicine, 40*(4), 434-439.

Blangiardo, M., Cameletti, M., Baio, G., & Rue, H. (2013). Spatial and spatio-temporal models with R-INLA. *Spatial and spatio-temporal epidemiology, 7*, 39-55.

Briggs, R. (2012). Spatial Autocorrelation concepts. Retrived from http//www.utdallas.edu/~briggs/.

Burtis, J. C., Sullivan, P., Levi, T., Oggenfuss, K., Fahey, T. J., & Ostfeld, R. S. (2016). The impact of temperature and precipitation on blacklegged tick activity and Lyme disease incidence in endemic and emerging regions. *Parasites & Vectors, 9*, 606. doi:10.1186/s13071-016-1894-6

Cliff, A., & Ord, J. (1975). The choice of a test for spatial autocorrelation. *Display and analysis of spatial data*, 54-77.

Cliff, A. D., & Ord, J. K. (1981). *Spatial processes: models & applications* (Vol. 44): Pion London.

Cressie, N. (2015). *Statistics for spatial data*: John Wiley & Sons.

Dijkstra, A., Janssen, F., De Bakker, M., Bos, J., Lub, R., Van Wissen, L. J., & Hak, E. (2013). Using spatial analysis to predict health care use at the local level: a case study of type 2 diabetes medication use and its association with demograpHic change and socioeconomic status. *PLoS One*, *8*(8), e72730.

Eisen, R. J., Eisen, L., & Beard, C. B. (2016). County-Scale Distribution of Ixodes scapularis and Ixodes pacificus (Acari: Ixodidae) in the Continental United States. *J Med Entomol, 53*(2), 349-386. doi:10.1093/jme/tjv237

Engelgau, M. M., Geiss, L. S., Saaddine, J. B., & et al. (2004). THe evolving diabetes burden in the united states. *Annals of Internal Medicine*, *140*(11), 945-950. doi:10.7326/0003-4819-140-11-200406010-00035

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2003). *Geographically weighted regression: the analysis of spatially varying relationships*: John Wiley & Sons.

Gelfand, A. E., Diggle, P., Guttorp, P., & Fuentes, M. (2010). *Handbook of spatial statistics*: CRC press.

Haire-Joshu, D. L. (2015). Next Steps: Eliminating Disparities in Diabetes and Obesity. *Preventing chronic disease*, *12*.

Hanrahan, J. P., Benach, J. L., Coleman, J. L., Bosler, E. M., Morse, D. L., Cameron, D. J., . . . Kaslow, R. A. (1984). Incidence and cumulative frequency of endemic Lyme disease in a community. *Journal of Infectious Diseases*, *150*(4), 489-496.

Hipp, J. A., & Chalise, N. (2015). Peer Reviewed: Spatial Analysis and Correlates of County-Level Diabetes Prevalence, 2009–2010. *Preventing chronic disease*, *12*.

Jones, C. J., & Kitron, U. D. (2000). Populations of Ixodes scapularis (Acari: Ixodidae) are modulated by drought at a Lyme disease focus in Illinois. *J Med Entomol, 37*(3), 408-415. doi:10.1603/0022-2585(2000)037[0408:poisai]2.0.co;2

Kauhl, B., Heil, J., Hoebe, C. J., Schweikart, J., Krafft, T., & Dukers-Muijrers, N. H. (2015). The spatial distribution of hepatitis C virus infections and associated determinants—an application of a geographically weighted poisson regression for evidence-based screening interventions in hotspots. *PLoS One, 10*. doi:10.1371/journal.pone.0135656

Killilea, M. E., Swei, A., Lane, R. S., Briggs, C. J., & Ostfeld, R. S. (2008). Spatial dynamics of Lyme disease: a review. *EcoHealth, 5*(2), 167-195.

Lawson, A. B. (2013). *Bayesian disease mapping: hierarchical modeling in spatial epidemiology*: CRC press.

Li, J., Kolivras, K. N., Hong, Y., Duan, Y., Seukep, S. E., Prisley, S. P., . . . Gaines, D. N. (2014). Spatial and Temporal Emergence Pattern of Lyme Disease in Virginia. *The American Journal of Tropical Medicine and Hygiene, 91*(6), 1166-1172. doi:10.4269/ajtmh.13-0733

Maes, E., Lecomte, P., & Ray, N. (1998). A cost-of-illness study of Lyme disease in the United States. *Clin Ther, 20*(5), 993-1008; discussion 1992.

Musenge, E., Chirwa, T. F., Kahn, K., & Vounatsou, P. (2013). Bayesian analysis of zero inflated spatiotemporal HIV/TB child mortality data through the INLA and SPDE approaches: applied to data observed between 1992 and 2010 in rural North East South Africa. *International Journal of Applied Earth Observation and Geoinformation, 22*, 86-98.

Ogden, N. H., St-Onge, L., Barker, I. K., Brazeau, S., Bigras-Poulin, M., Charron, D. F., . . . Maarouf, A. (2008). Risk maps for range expansion of the Lyme disease vector, Ixodes scapularis, in Canada now and with climate change. *International Journal of Health Geographics, 7*(1), 24.

Orloski, K. A., Hayes, E. B., Campbell, G. L., & Dennis, D. T. (2000). Surveillance for Lyme disease—United States, 1992–1998. *MMWR CDC Surveill Summ, 49*(3), 1-11.

Plant, R. E. (2012). *Spatial data analysis in ecology and agriculture using R*: CRC Press.

Raghavan, R. K., Goodin, D. G., Neises, D., Anderson, G. A., & Ganta, R. R. (2016). Hierarchical Bayesian Spatio–Temporal Analysis of Climatic and Socio–Economic Determinants of Rocky Mountain Spotted Fever. *PLoS One, 11*(3), e0150180.

Ripley, B. D. (2005). *Spatial statistics* (Vol. 575): John Wiley & Sons.

Schauber, E. M., & Ostfeld, R. S. (2002). Modeling the effects of reservoir competence decay and demographic turnover in Lyme disease ecology. *Ecological Applications, 12*(4), 1142-1162.

Shaw, J. E., Sicree, R. A., & Zimmet, P. Z. (2010). Global estimates of the prevalence of diabetes for 2010 and 2030. *Diabetes research and clinical practice, 87*(1), 4-14.

Steere, A. C., Coburn, J., & Glickstein, L. (2004). The emergence of Lyme disease. *The Journal of clinical investigation, 113*(8), 1093-1101.

Steere, A. C., Hardin, J. A., & Malawista, S. E. (1978). Lyme arthritis: a new clinical entity. *Hosp Pract, 13*(4), 143-158.

Steere, A. C., Taylor, E., Wilson, M. L., Levine, J. F., & Spielman, A. (1986). Longitudinal assessment of the clinical and epidemiological features of Lyme disease in a defined population. *Journal of Infectious Diseases, 154*(2), 295-300.

Subak, S. (2003). Effects of climate on variability in Lyme disease incidence in the northeastern United States. *American Journal of Epidemiology*, *157*(6), 531-538.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic geography*, 234-240.

Wild, S., Roglic, G., Green, A., Sicree, R., & King, H. (2004). Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. *Diabetes Care, 27* (10): 2568-2569.

Wang, X., Chen, M.-H., Kuo, R. C., & Dey, D. K. (2015). BAYESIAN SPATIAL-TEMPORAL MODELING OF ECOLOGICAL ZERO-INFLATED COUNT DATA. *Statistica Sinica, 25*(1), 189-204. doi:10.5705/ss.2013.212w

Chapter 2

2 Spatial Analysis of Diabetes Prevalence in the Midwestern United States using Mixed Geographically Weighted Regression Models

ABSTRACT

Diabetes is a major health problem in the United States. There is an increasing interest in the relationship between diabetes and sociodemographic and lifestyle factors but the extent of the geographical variability of diabetes with respect to these variables still remains unclear. The regression models commonly used for disease modeling either use Ordinary Least Square (OLS) regression by assuming all the explanatory variables have the same effect over geographical locations or Geographically Weighted Regression (GWR) that assumes the effect of all the explanatory variables vary over the geographical space. In reality, the effect of some of the variables may be fixed (global) and other variables vary spatially (local). For this type of data analysis, Mixed Geographically Weighted Regression (MGWR) which can include global and local variables in the same model is the best alternative (Fotheringham et al., 2003). We propose using MGWR model to study the association between diabetes prevalence rate and sociodemographic and life style variables in counties of the Midwestern United States. The results of this study showed that the effect of some of the variables is global and others are local. The benefit of fitting MGWR is it gives local insight of the disease, which helps policy makers develop effective policy to address disease at the local level.

Keywords: Geographically Weighted regression Model (GWR), Mixed Geographically Weighted regression Model (MGWR), Ordinary Least Square (OLS) regression, Diabetes prevalence rate.

2.1 Introduction

Diabetes is a serious health problem in the Unites States. According to CDC report, more than 29 million people had diabetes with estimated total health care cost of $ 245 billion in 2012. Type 2 diabetes that accounts for more than 90 percent of total diabetes affects people of all sex, age, race and ethnic groups; however, the rate is higher in American Indians, African Americans and people with socioeconomic disadvantages (Haire-Joshu, 2015). The findings of the previous studies have shown that diabetes is associated with increased risk of microvascular complications (Klein, 1995; Pirart, 1978), myocardial infractions (Kuusisto, Mykkänen, Pyörälä, & Laakso, 1994; Turner et al., 1998), stroke (Lehto, Rönnemaa, Pyörälä, & Laakso, 1996). Diabetes is associated with obesity, physical inactivity, race and some other socioeconomic covariates (Hipp & Chalise, 2015). There is a steady increase in type 2 diabetes prevalence especially in adolescents and African Americans (Arslanian, 2000; Arslanian, Bacha, Saad, & Gungor, 2005; Harris, 2001).

Studies of the correlates of diabetes ignore the spatial non-stationarity by either fitting OLS method or using all the variables as nonstationary by fitting GWR model. A number of studies (Chen, Wu, Yang, & Su, 2010; Dijkstra et al., 2013; Hipp & Chalise, 2015; Siordia, Saenz, & Tom, 2012) used GWR model to study the association between diabetes and other covariates.

GWR is one of the localized regression techniques which accounts for spatial heterogeneity or spatial non- stationarity (Benson, Chamberlin, & Rhinehart, 2005; C. Brunsdon, Fotheringham, & Charlton, 1996; Fotheringham, Brunsdon, & Charlton, 2003; Lu, Harris, Charlton, & Brunsdon, 2015). As an exploratory tool, GWR is useful in wide varieties of research fields including but not limited to health and disease (Chalkias et al., 2013; Chen et al., 2010; Chi, Grigsby-Toussaint, Bradford, & Choi, 2013; Dijkstra et al., 2013; Fraser, Clarke, Cade, & Edwards, 2012; Hipp & Chalise, 2015; Lin & Wen, 2011; Nakaya, Fotheringham, Brunsdon, & Charlton, 2005; Schuurman, Peters, & Oliver, 2009; Siordia et al., 2012; Wen, Chen, & Tsai, 2010; Yang, Wu, Chen, & Su, 2009), housing market (Fotheringham et al., 2003; Yu, Wei, & Wu, 2007), poverty (Benson et al., 2005; Farrow, Larrea, Hyman, & Lema, 2005; Longley & Tobón, 2004), traffic models (Selby & Kockelman, 2013; Zhao & Park, 2004), forest fire (Martínez-Fernández, Chuvieco, & Koutsias, 2013; Mitchell & Yuan, 2010; Sá et al., 2011), crime (Cahill & Mulligan, 2007; Troy, Grove, & O'Neil-Dunne, 2012; Wheeler & Waller, 2009; Yan, Shu, & Yuan, 2010; Haifeng Zhang & Song, 2014), fisheries and wildlife (Irigoien et al., 2014; Sheehan, Strager, & Welsh, 2013; Tseng et al., 2013; Windle, Rose, Devillers, & Fortin, 2009), and tourism (Deller, 2010; Honglei Zhang, Zhang, Lu, Cheng, & Zhang, 2011).

One should not expect the effect of every explanatory variable always significantly vary spatially. If this is the case, then the use of GWR which considers every explanatory variable significantly varies over space leads to inefficient or incorrect conclusions (Kang & Dall'erba, 2016; Wei & Qi, 2012). Most of the studies that report GWR consider all the coefficients vary spatially without testing whether the spatial differences are statistically significant.

MGWR model can include both global and local variables in a single model. The use of MGWR is efficient and easy to apply. There are a large number of studies that consider GWR model in disease epidemiology. The use of MGWR model has been ignored in the studies of disease epidemiology. To the best of our knowledge, none of the studies have used MGWR model before for fitting spatial regression model for any epidemiological data. The objective of this study is to find the relationship between sociodemographic and lifestyle factors in the geographical variability of diabetes in the Midwestern United States by using MGWR.

2.2 Methods

2.2.1 Geographically Weighted Regression

A global regression model (OLS) can be written as:

$$y_i = \beta_0 + \sum_k \beta_k x_{ik} + \varepsilon_i \ldots \ldots \ldots (1)$$

GWR is an extended version of traditional regression estimates of local rather than global parameters. GWR model can be written as:

$$y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i) x_{ik} + \varepsilon_i \ldots \ldots \ldots (2)$$

Where (u_i,v_i) represents the coordinates of the i^{th} point in the space and $\beta_k(u_i,v_i)$ is a realized value of the continuous function $\beta_k(u,v)$ at point i. GWR model allows the continuous surface of parameter values and the measured value at certain points denote the spatial variability of the surface.

From equation (2), we can assume that the near observations to the location i are more influential to the estimates of β_k (u_i,v_i) than observations farther from location i. GWR model is based on the weight of an observation based on the proximity to the location i. More weight is given to the data with observations that are close to i than data with observations farther away.

$$\hat{\boldsymbol{\beta}}(u_i, v_i) = (\mathbf{X}^T \mathbf{W}(u_i, v_i)\mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}(u_i, v_i)\mathbf{y} \ldots \ldots \ldots (3)$$

Where $\hat{\boldsymbol{\beta}}$ is the estimated value of $\boldsymbol{\beta}$, and $\mathbf{W}(u_i, v_i)$ is an n by n matrix with off-diagonal elements are zero and diagonal elements are geographical weights of each of the n observed data for regression point i.

To see this more clearly, the classical regression equation can be written in matrix form as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \ldots \ldots \ldots (4)$$

Where $\boldsymbol{\beta}$ is the vector of parameters to be estimated, which is constant over space and estimated by:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \ldots \ldots \ldots (5)$$

The GWR equivalent of this model is:

$$\mathbf{Y} = (\boldsymbol{\beta} \otimes \mathbf{X})1 + \boldsymbol{\varepsilon} \ldots \ldots \ldots (6)$$

The matrix $\boldsymbol{\beta}$, which has n sets of local parameters. It has the following structures:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0(u_1,v_1)\ \beta_1(u_1,v_1)\ldots\beta_k(u_1,v_1) \\ \beta_0(u_2,v_2)\ \beta_1(u_2,v_2)\ldots\beta_k(u_2,v_2) \\ \ldots \qquad\quad \ldots \qquad\quad \ldots \\ \beta_0(u_n,v_n)\ \beta_1(u_n,v_n)\ldots\beta_k(u_n,v_n) \end{bmatrix} \ldots\ldots\ldots (7)$$

The parameter estimates of each row of the above matrix is given by:

$$\hat{\boldsymbol{\beta}}(i) = (\mathbf{X}^T\mathbf{W}(i)\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}(i)\mathbf{Y} \ldots\ldots\ldots (8)$$

Where i represents a row of the matrix in (7) and **W** (i) is an n by n spatial weighting matrix of the form:

$$\mathbf{W}(i) = \begin{bmatrix} w_{i1} & 0 & \ldots\ldots\ldots & 0 \\ 0 & w_{i2} & \ldots\ldots\ldots & 0 \\ . & . & & . \\ . & . & & . \\ 0 & 0 & \ldots\ldots\ldots & w_{in} \end{bmatrix} \ldots\ldots\ldots (9)$$

Where w_{in} denotes the weight given to the data point n in the calibration of model for location i.

The implicit weighting scheme for OLS framework is:

$$w_{ij} = 1 \quad \forall i, j$$

Where j is a specific point in space at which data are observed and i represents any points in space for which parameters are estimated. The global model has a weight of unity. The initial step of weighting based local model excludes the observations outside some distance b from regression point. The weighting function can be written as:

$$w_{ij} = 1\, if\ d_{ij} < b$$
$$w_{ij} = 0\, otherwise$$

Spatial weighting has the problem of discontinuity, which results in a drastic change in estimated coefficients. One way of overcoming the problem of discontinuity is to express w_{ij} as a continuous function of d_{ij}, the distance between i and j as:

$$w_{ij} = \exp[-\frac{1}{2}(d_{ij}/b)^2]$$

Where b is referred as bandwidth. An alternative kernel utilizes the bi-square function as:

$$w_{ij} = [1-(d_{ij}/b)^2]^2 \quad if\ d_{ij} < b$$
$$= 0 \quad otherwise$$

This will provide a continuous weighting function up to distance b and then zero weight for any data point outside b.

Bandwidth Selection:

These methods determine the optimum bandwidth in GWR:

Least cross-validation score (CV):

Cross-validation score is the difference between observed value and the GWR calibrated value using the bandwidth.

$$CV = \sum_i [y_i - \hat{y}_{\neq i}(b)]^2$$

Where $\hat{y}_{\neq i}(b)$ is the fitted value of y_i with data from point i is omitted from the calibration. The lower value of CV indicates better model fit.

Least Akaike Information Criterion (AIC):

Akaike Information Criterion (AIC) derives the optimum bandwidth for GWR as:

$$AICc = 2n\log_e(\hat{\sigma}) + n\log_e(2\pi) + n\left\{\frac{n + tr(\mathbf{S})}{n - 2 - tr(\mathbf{S})}\right\}$$

Where n is the sample size, $\hat{\sigma}$ is the estimated standard deviation of the error term and tr(**S**) is the trace of the hat matrix. $\hat{\mathbf{y}} = \mathbf{Sy}$, where **S** is hat matrix which maps fitted values on to observed values. Hat matrix is a function of the bandwidth of weighting function.

Each row of **S**, $\mathbf{r}_i$ is given by:

$$\mathbf{r}_i = \mathbf{X}_i(\mathbf{X}^T\mathbf{W}(u_i,v_i)\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}(u_i,v_i)$$

We used cross validation score method to select optimum bandwidth.

Model selection

The best model selection was done by the following algorithm as described by Gollini et al. (2013):

(1) All the possible bivariate GWR model was calibrated by sequentially regressing a single independent variable against the dependent variable

(2) The best performing model with least AICc value, and permanently include the independent variable in the subsequent models.

(3) Remaining independent variables were sequentially introduced to construct the new model with permanently included independent variables. The next permanently included variable is selected based on AICc value

(4) Step 3 was repeated until all the independent variable were included in the model.

2.2.2 Mixed Geographically Weighted Regression (MGWR) model

MGWR is an extension of basic GWR model when the degree of variation for some of the coefficients might be negligible. MGWR model has two different types of coefficients. Some of the coefficients are global and the others are local. Global coefficient has fixed effect over space whereas local coefficients are modeled as the function of geographical locations (Benson et al., 2005; Chris Brunsdon, Fotheringham, & Charlton, 2000; Fotheringham et al., 2003; Mei, He, & Fang, 2004). According to Fotheringham et al.(2003) MGWR model can be written as:

$$y_i = \sum_{j=1,k_a} a_j x_{ij}(a) + \sum_{l=1,k_b} b_l(u_i, v_i) x_{il}(b) + \varepsilon_i$$

$$\mathbf{X_a} = \begin{pmatrix} x_{11} & x_{12} & . & . & . & x_{1q} \\ x_{21} & x_{22} & . & . & . & x_{2q} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ x_{n1} & x_{n2} & . & . & . & x_{nq} \end{pmatrix}., \ \mathbf{X_b} = \begin{pmatrix} x_{1,q+1} & x_{1,q+2} & . & . & . & x_{1,p} \\ x_{2,q+1} & x_{2,q+2} & . & . & . & x_{2,p} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ x_{n,q+1} & x_{n,q+2} & . & . & . & x_{n,p} \end{pmatrix}, \mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{pmatrix}$$

$$\boldsymbol{\beta}_\mathbf{a} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_q \end{pmatrix}, \boldsymbol{\beta}_\mathbf{b}(u_i, v_i) = \begin{pmatrix} \beta_{q+1}(u_i, v_i) \\ \beta_{q+2}(u_i, v_i) \\ . \\ . \\ . \\ \beta_p(u_i, v_i) \end{pmatrix} \text{ where i= 1, 2, n,}$$

$$\mathbf{S_a} = \mathbf{X_a}\,(\mathbf{X_a}^{\mathbf{T}}\mathbf{X_a})^{-1}\mathbf{X_a}^{\mathbf{T}},\ \mathbf{S_b} = \begin{pmatrix} \mathbf{X_{b1}}^{\mathbf{T}}[\mathbf{X_b}^{\mathbf{T}}\mathbf{W}(u_1,v_1)\mathbf{X_b}]^{-1}\mathbf{X_b}^{\mathbf{T}}\mathbf{W}(u_1,v_1) \\ \mathbf{X_{b2}}^{\mathbf{T}}[\mathbf{X_b}^{\mathbf{T}}\mathbf{W}(u_1,v_2)\mathbf{X_b}]^{-1}\mathbf{X_b}^{\mathbf{T}}\mathbf{W}(u_2,v_2) \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{X_{bn}}^{\mathbf{T}}[\mathbf{X_b}^{\mathbf{T}}\mathbf{W}(u_n,v_n)\mathbf{X_b}]^{-1}\mathbf{X_b}^{\mathbf{T}}\mathbf{W}(u_n,v_n) \end{pmatrix}$$

Where

$\mathbf{X^T}_{bi} = (X_{i,\,q+1},\ X_{i,\,q+2},\ldots,X_{i,\,p})$ is the i^{th} row of the $\mathbf{X_b}$ and

$\mathbf{W}(u_i,v_i)$ = diag $[w_1(u_i,v_i),\ w_2(u_i,v_i),\ldots\ w_n(u_i,v_i)]$ is an n x n diagonal weight matrix at location (u_i,v_i) where elements in diagonal are usually taken to be a Gaussian function of the form :

$$W_j(u_i,v_i) = \exp\left[-\left(\frac{d_{ij}}{b}\right)^2\right],\ \text{j=1, 2, … , n}$$

where b is bandwidth.

Calibration approach in (Fotheringham et al. 2002, chapter 3):

$$\hat{\boldsymbol{\beta}}_{\mathbf{a}} = \left(\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_q\right)^T = \left[\mathbf{X_a}^{\mathbf{T}}\left(\mathbf{I}-\mathbf{S_a}\right)^{\mathbf{T}}\left(\mathbf{I}-\mathbf{S_a}\right)\mathbf{X_a}\right]^{-1}\mathbf{X_a}^{\mathbf{T}}\left(\mathbf{I}-\mathbf{S}_b\right)^{\mathbf{T}}\left(\mathbf{I}-\mathbf{S}_b\right)\mathbf{Y},$$

and spatially varying coefficient vector at location (u_i,v_i) as:

$$\hat{\boldsymbol{\beta}}_{\mathbf{b}}(u_i,v_i) = \left[\hat{\beta}_{q+1}(u_i,v_i), \hat{\beta}_{q+2}(u_i,v_i), \cdots, \hat{\beta}_p(u_i,v_i)\right]^T$$

$$= \left[\mathbf{X_b}^{\mathbf{T}}\mathbf{W}(u_i,v_i)\mathbf{X_b}\right]^{-1}\mathbf{X_b}\,\mathbf{W}(u_i,v_i)\left(\mathbf{Y}-\mathbf{X_a}\hat{\boldsymbol{\beta}}_{\mathbf{a}}\right)\ \text{i= 1, 2, … , n.}$$

Therefore, the fitted values of the response at n locations are obtained by:

$$\hat{\mathbf{Y}} = (y_1, y_2, \cdots, y_n)^T = \mathbf{S_b}\left(\mathbf{Y} - \mathbf{X_a}\hat{\boldsymbol{\beta}}_\mathbf{a}\right) + \mathbf{X_a}\hat{\boldsymbol{\beta}}_\mathbf{a} = \mathbf{S_b Y} + (\mathbf{I} - \mathbf{S_b})\mathbf{X_a}\hat{\boldsymbol{\beta}}_a = \mathbf{SY}$$

Where,

$$\mathbf{S} = \mathbf{S_b} + (\mathbf{I} - \mathbf{S_b})\mathbf{X_a}\left[\mathbf{X_a}^{\mathbf{T}}(\mathbf{I} - \mathbf{S_b})\mathbf{X_a}\right]^{-1}\mathbf{X_a}^{\mathbf{T}}(\mathbf{I} - \mathbf{S_b})^{\mathbf{T}}(\mathbf{I} - \mathbf{S_b})$$

Where for each observation i, y_i is the dependent variable, (u_i, v_i) represents geographical location, k_a represents global coefficients and k_b represents local coefficients. The group of independent variables associated with global coefficients is referred as a-group variables and independent variables associated with local coefficients are referred as b-group variables. There is one intercept term from either a-group or b-group of variables but not for both.

2.2.3 Data Source:

This study includes the county-level data for the Midwestern United States. It included county-level data of 1055 counties from Midwestern States, Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, Ohio, North Dakota, South Dakota, and Wisconsin. This study includes data on diabetes, obesity rates, and physical inactivity for 2012 from the Centers for Disease Control and Prevention's Diabetes Interactive Atlas. The data collection is based on Behavioral Risk Factor Surveillance System (BRFSS). CDC defines diabetes prevalence as the estimated percentage of adults with either type 1 or type 2 diabetes after adjustment for age. Obesity prevalence is defined as the estimated percentage of adults with body mass index ≥ 30 after adjustment for age; physical inactivity prevalence is defined as percentage of adults who have not done any physical exercise or activity for past 30 days. Data for the socioeconomic variables –percentage nonwhite population, percentage living behind federal poverty level, percentage below high school

graduates, percentage unemployed, percent of adults in work force, and percent of people with German ancestry were collected from the US Census Bureau's American Community Survey 5-year estimates (2008-2012).

Percentage of people who did not identify himself or herself as white is referred as percentage of nonwhite population. The percentage of people living below federal poverty levels is determined based on income threshold defined by the United States census bureau, which varies depending on family size. Unemployment rate was defined as percentage of people aged 16 years or older that did not go for work for the reference week. The education variable is determined as the percentage of people who reported having less than a high school diploma. Percentage of people in labor force is determined as percentage of population that is either working or actively seeking employment. Percentage of German ancestry is defined as people who have defined their ancestry as German.

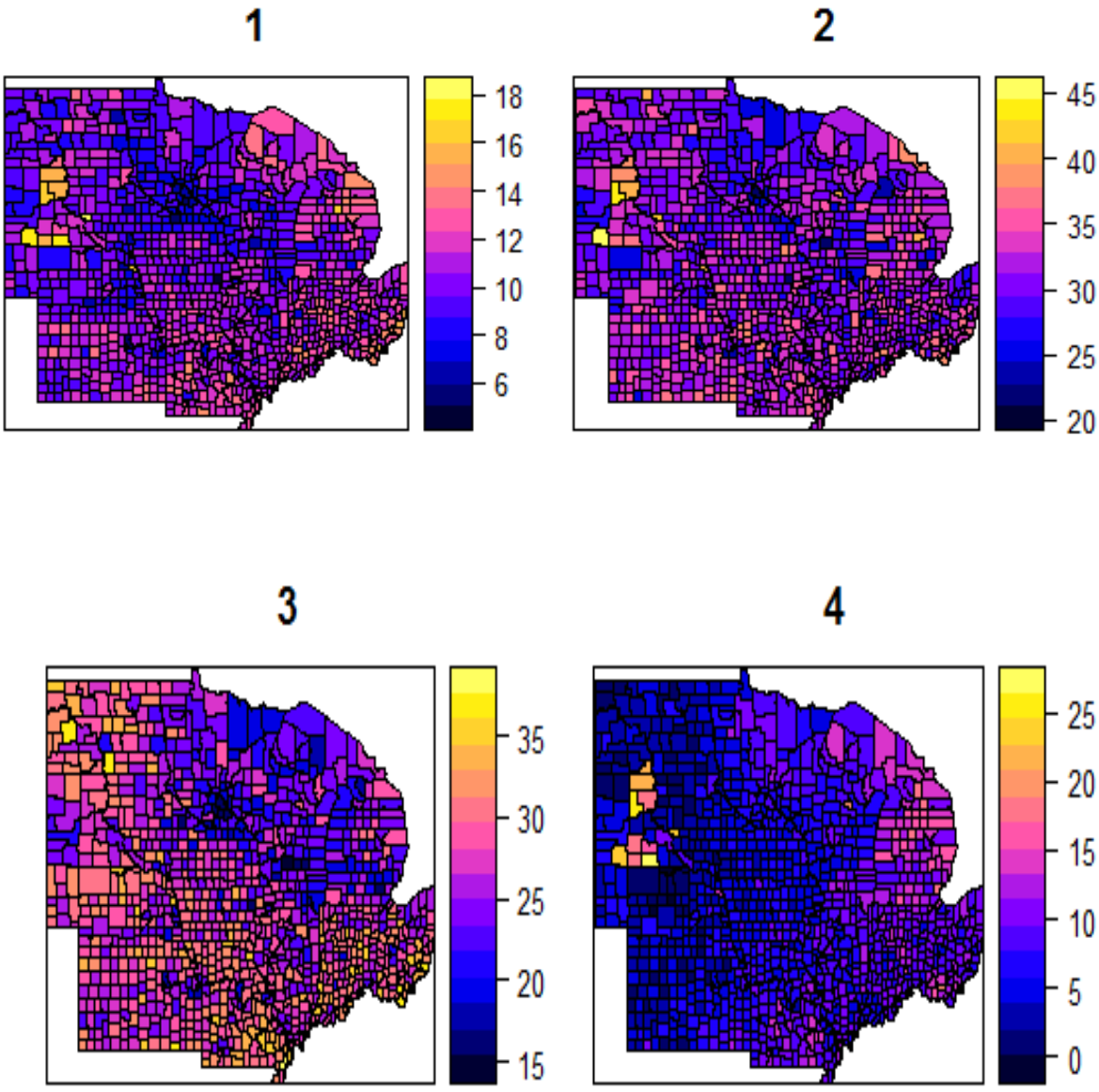
Table 2.1 presents the list of variables used in different models where diabetes is a dependent variable and the other variables are the independent variables.

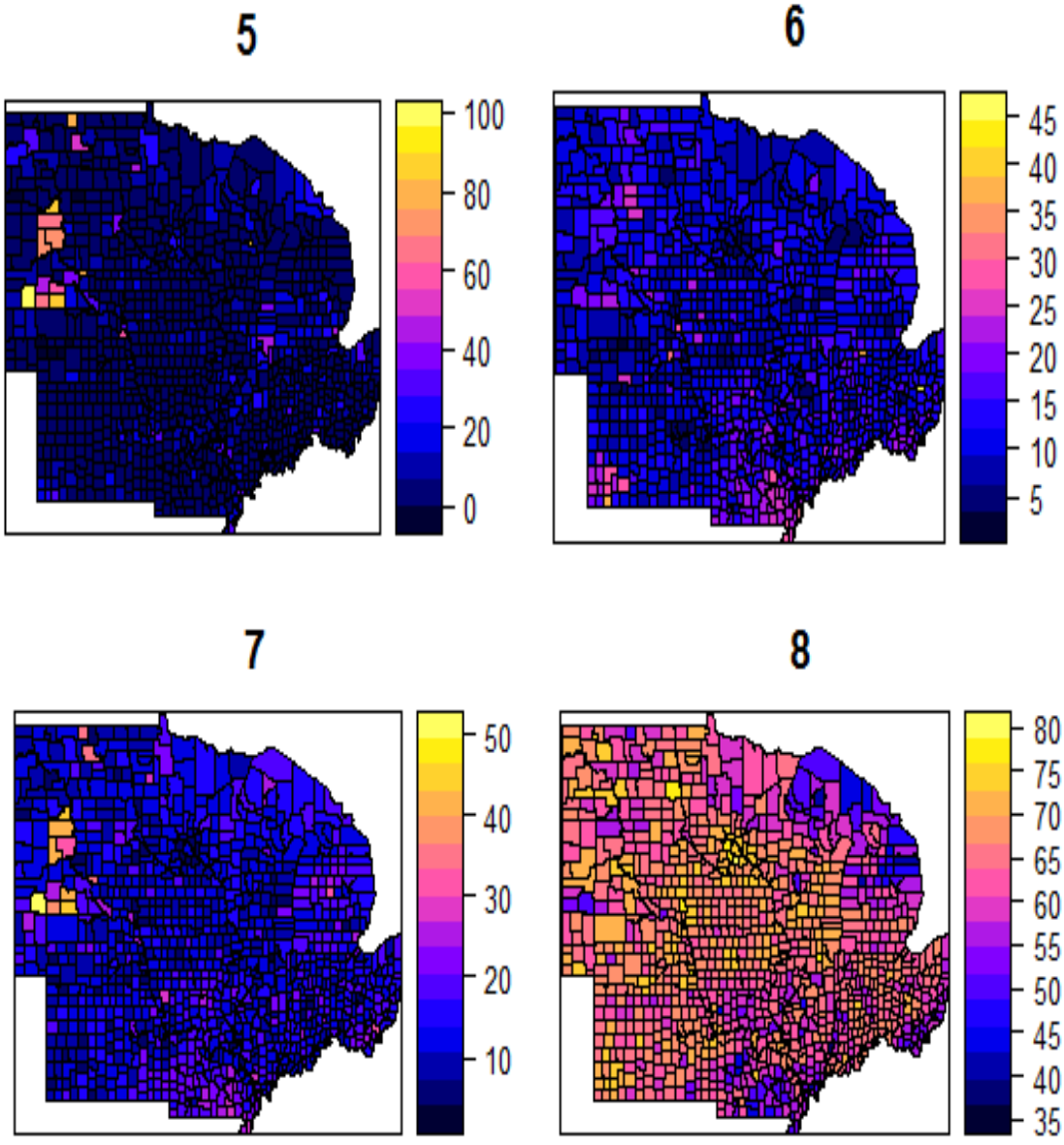
Table 2. 1 The list of response and explanatory variables used in model fitting.

Variables	Description of Variables	Variable status
Diabetes	Percentage of people with diabetes	Response
Obesity	Percentage of people with obesity	Explanatory
Physical inactivity	Percentage physically inactive	Explanatory
Unemployment	Percentage unemployed	Explanatory
Nonwhite	Percentage nonwhite population	Explanatory
Poverty	Percentage living below federal poverty level	Explanatory
Education	Percentage adults with less than high school diploma	Explanatory
Labor force	Percentage of people in labor force	Explanatory
German ancestry	Percentage of people with German ancestry	Explanatory

2. 3 Results

Figure 2.1 displays the distribution of diabetes and other variables in the Midwestern United States. It shows that there was variation in county-level data.




5
100
80
60
40
20
0
6
45
40
35
30
25
20
15
10
5
7
50
40
30
20
10
8
80
75
70
65
60
55
50
45
40
35

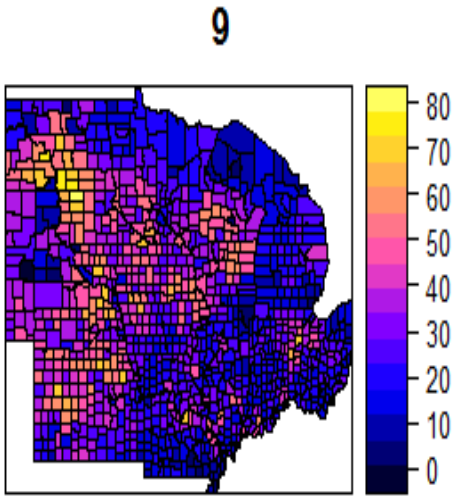


Figure 2. 1 Spatial distribution of variables: (1) Diabetes (2) Obesity (3) Physical inactivity (4) Unemployment (5) Nonwhites (6) Education (7) Poverty (8) Labor force (9) German ancestry.

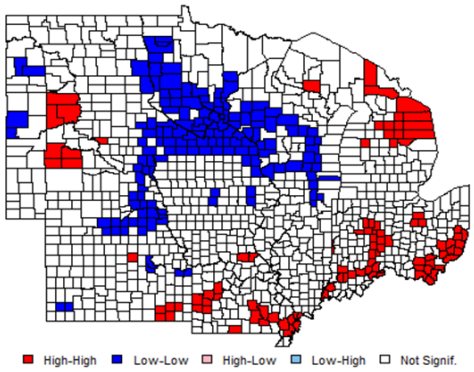
Table 2.2 presents the summary statistics of variables used for modeling. The minimum diabetes prevalence rate was 5 percent, median 10.7 percent, mean 10.66 percent and maximum 17.8 percent. Obesity rate was minimum 21 percent, median and mean both 31.5 percent and maximum 44.5 percent. Physical inactivity rate was minimum 15.5 percent, median 27.2 percent, mean 27.09 and maximum 37.6 percent. Unemployment rate was minimum 0 percent, median 6.8 percent, mean 7.22 percent, and maximum 26.4. Percent of adults with below high school diploma was minimum 3.2, median 11.4, mean 12.15, and maximum 44.5 percent. Percent of population below poverty level was minimum 3.9 percent, median 13 percent, mean 13.78, and maximum 49.5, percent of adults in labor force was minimum 36.4 percent, median 64.6 percent, mean 63.64 percent, and maximum 78.7 percent, and percent of nonwhite population was minimum 0 percent, median 2.9 percent, mean 6.18 percent, and maximum 96.2 percent.

Table 2. 2 Summary statistics of variables.

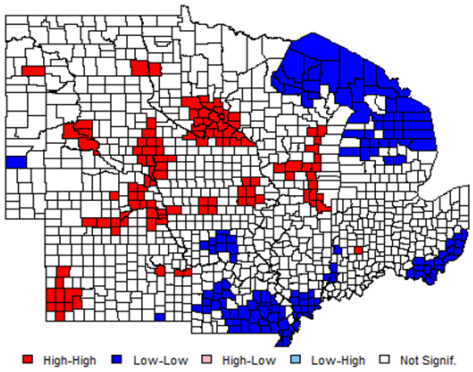
Variables	Minimum	Median	Mean	Maximum
Diabetes	5	10.7	10.66	17.8
Obesity	21	31.5	31.5	44.5
Physical Inactivity	15.5	27.2	27.09	37.6
Unemployment	0	6.8	7.22	26.4
Education	3.2	11.4	12.15	44.5
Poverty	3.9	13	13.78	49.5
Labor force	36.4	64.6	63.64	78.7
Nonwhite	0	2.9	6.18	96.2
German ancestry	0.31	28	30.55	77.93

Figure 2.2 shows that all of the variables have significant local spatial autocorrelation. All the areas with red color indicate the “hot spots” and areas with blue color indicate “cold spot” of the particular variable.

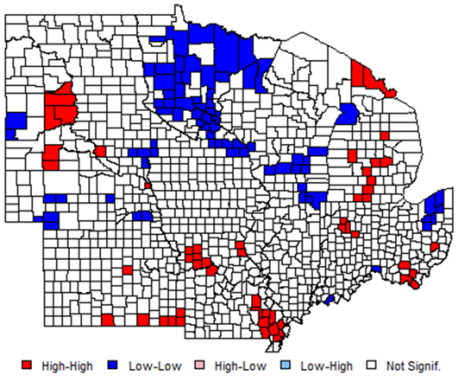
Local Moran's I (Diabetes Prevalance)



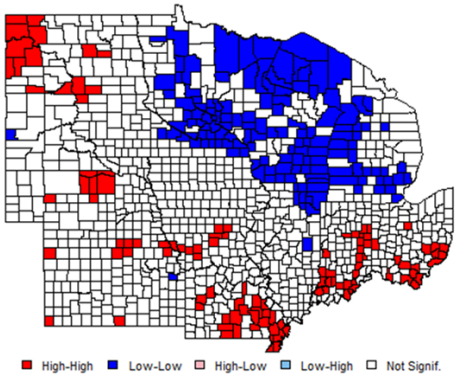
Local Moran's I (Percent in Labor Force)



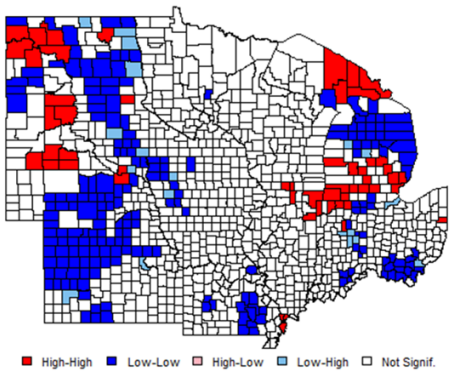
Local Moran's I (Obesity)



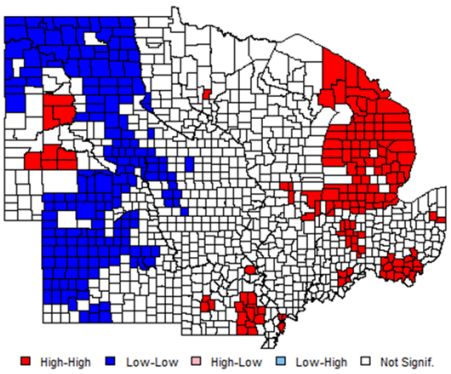
Local Moran's I (Physical Inactivity)



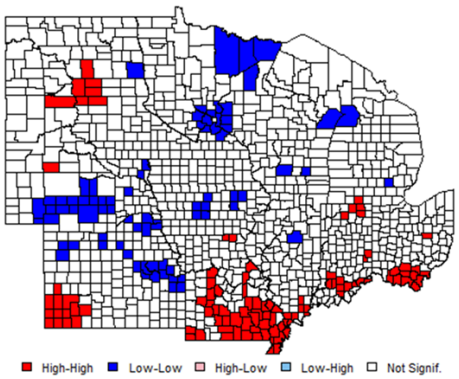
Local Moran's I (Percent of Non White Population



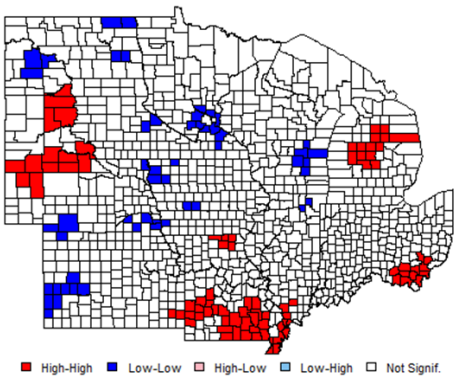
Local Moran's I (Unemployment Rate)



Local Moran's I (Below High School Education)



Local Moran's I (Poverty Rate)



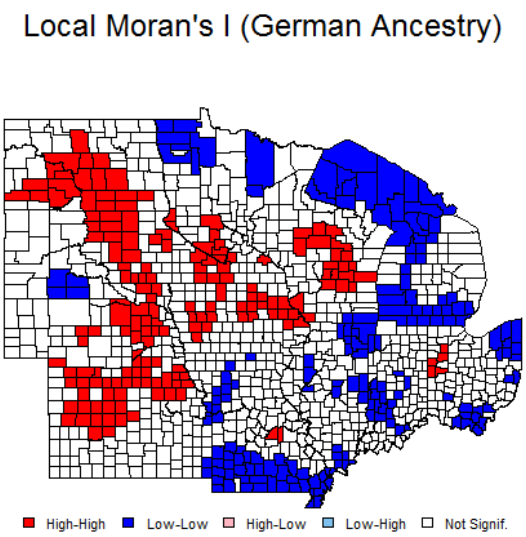


Figure 2. 2 Plots of local autocorrelation (Moran's I).

Table 2.3 presents the regression parameter estimates for the OLS model. The results from OLS model shows that all the variables except percent below high school, German ancestry, and percent below the poverty level are significant at 0.05 significance level. Obesity, physical inactivity, nonwhite population, and unemployment have positive relationship with diabetes prevalence and percent of people in labor force have a negative relationship with diabetes prevalence in the Midwestern counties in the United States. The Moran's I test for spatial autocorrelation for OLS residuals was significant that means the residuals of OLS model is spatially autocorrelated. The model coefficients of OLS model are not reliable because it violates the independence of the residuals in OLS model.

Table 2. 3 Coefficients of OLS model.

Variables	Estimates	Standard Error	t value	P-values
Intercept	7.07	0.73	9.71	<0.0001**
Obesity	0.15	0.01	10.75	<0.0001**
Physical inactivity	0.15	0.01	14.31	<0.0001**
Unemployment	0.13	0.01	9.52	<0.0001**
Nonwhite	0.01	0.004	2.32	0.02**
Poverty	-0.01	0.01	-1.29	0.2
Education	-0.01	0.009	-1.23	0.22
Labor force	-0.09	0.007	12.715	<0.0001**
German ancestry	0.001	0.002	0.44	0.66
AIC	3184.56			

** means those variables are significant at 0.05 level of significance

Table 2.4 and figure 2.5 present the parameter estimates for GWR model coefficients. The Moran's I test for spatial autocorrelation of residuals from the GWR model was not significant. This means that there was no issue of spatial autocorrelation in GWR residuals.

Table 2. 4 Coefficients for GWR model.

Variables	Minimum	First Quartile	Median	Third Quartile	Maximum
Intercepts	1.23	5.08	7.22	10.4	15.66
Obesity	0.05	0.11	0.15	0.18	0.22
Physical inactivity	0.02	0.1	0.13	0.15	0.22
Unemployment	0.002	0.04	0.06	0.09	0.17
Nonwhite	-0.05	-0.01	0.003	0.02	0.07
Poverty	-0.09	-0.04	-0.01	0.002	0.03
Education	-0.07	-0.02	0.01	0.03	0.06
Labor force	-0.17	-0.12	-0.09	-0.05	-0.006
German ancestry	-0.01	-0.005	0.0004	0.007	0.01
AICc	3063.92				

GWR model was used to model spatial autocorrelation. The selection of variables was done based on model AICs values. From figure 2.3 and 2.4, The GWR model with all the variables included in OLS model was the best model.

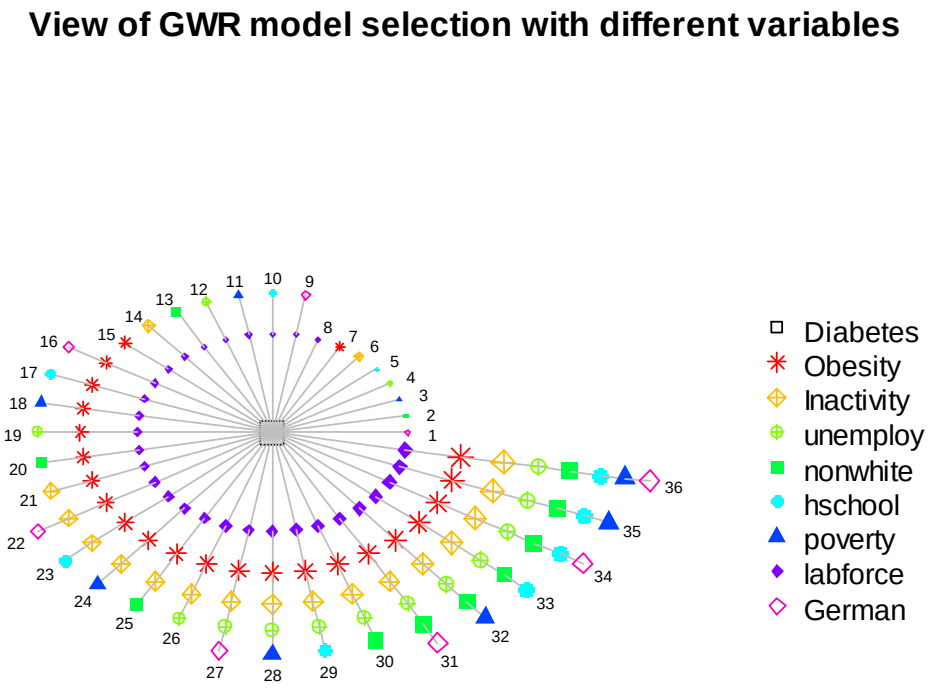


Figure 2. 3 View of GWR model selection with different variables.

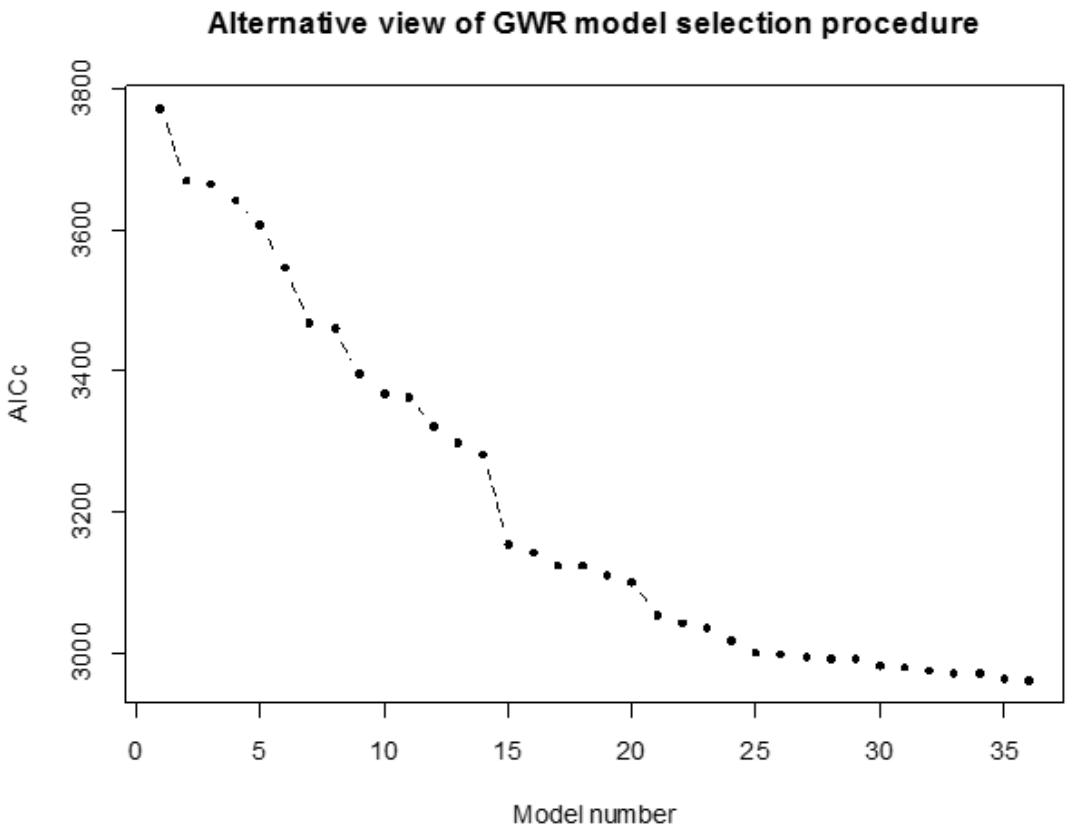
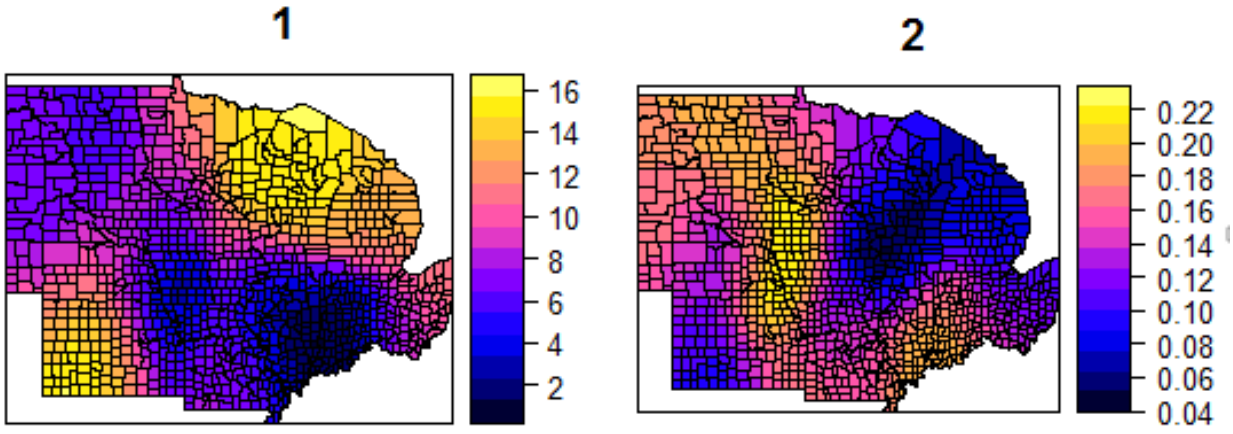
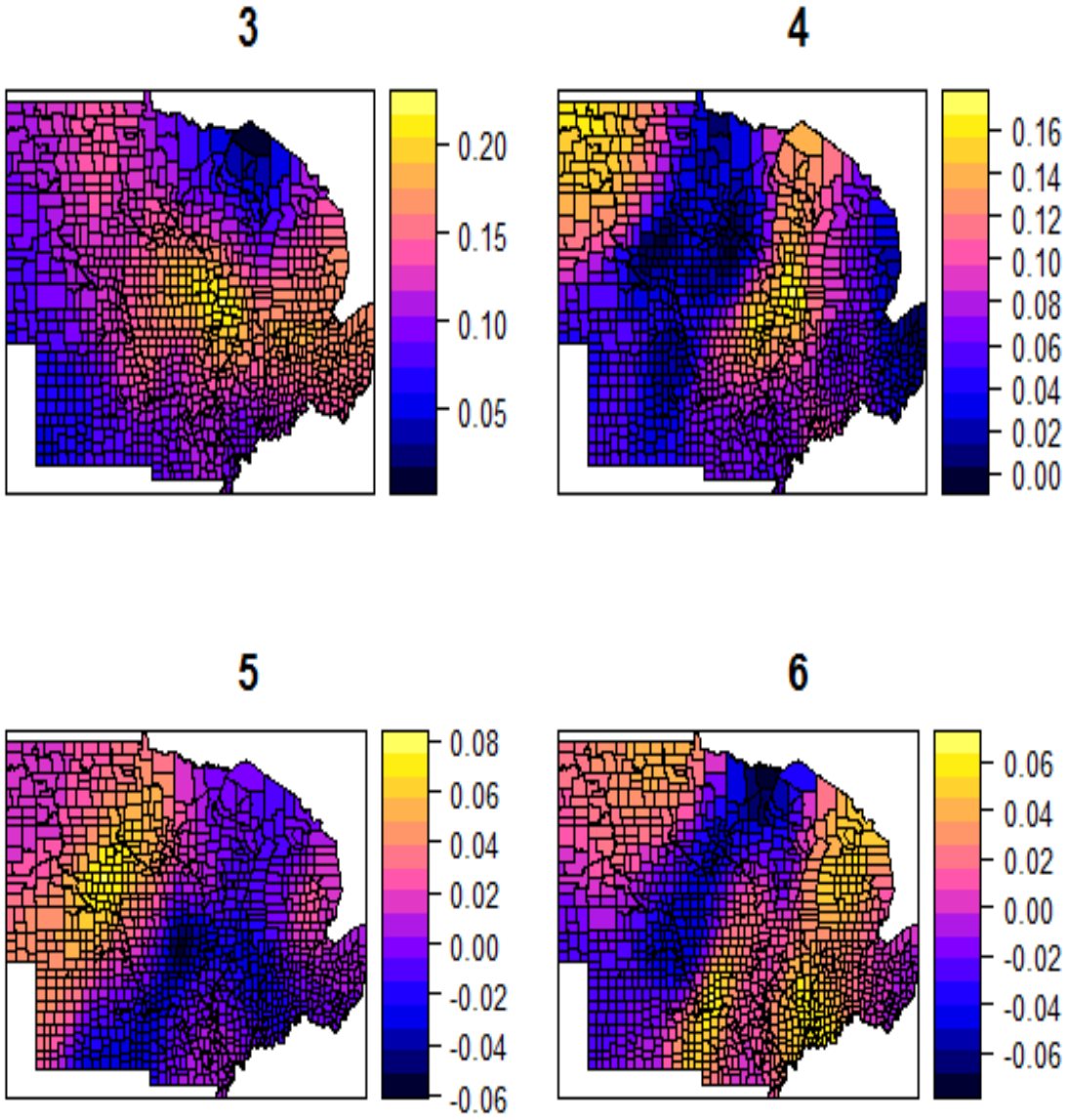


Figure 2. 4 Alternative view of GWR model selection procedure.




3
0.20
0.15
0.10
0.05
4
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
5
0.08
0.06
0.04
0.02
0.00
-0.02
-0.04
-0.06
6
0.06
0.04
0.02
0.00
-0.02
-0.04
-0.06

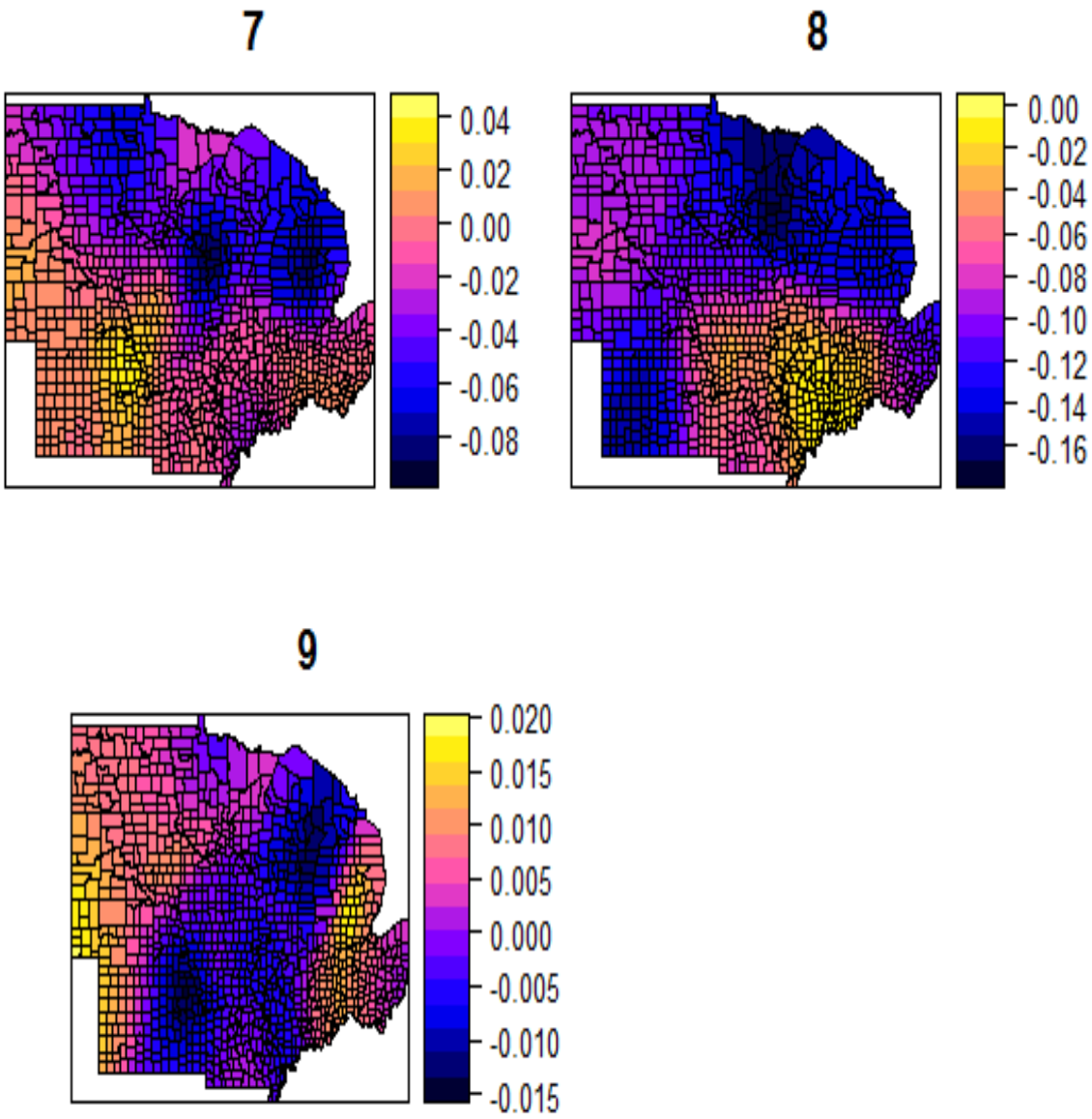


Figure 2.5 GWR coefficients: (1) Intercepts (2) Obesity (3) Physical inactivity (4) Unemployment (5) Nonwhites (6) Education (7) Poverty (8) Labor force (9) German ancestry.

The results in table 2.5 from Monte Carlo test for spatial non stationarity of variables show that obesity, physical inactivity, nonwhites, education, and labor force have local effect and unemployment, poverty and German ancestry has the global effect. Since poverty, unemployment, and German ancestry are stationarity variables.

Table 2. 5 Monte Carlo test of nonstationary of variables.

Variables	P value
Intercepts	0
Obesity	0.02
Physical inactivity	0.01
Unemployment	0.2
Nonwhite	0
Poverty	0.41
Education	0.001
Labor force	0
German ancestry	0.43

The MGWR model includes poverty, unemployment and German ancestry as global variables and obesity, physical inactivity, nonwhites, education, and labor force as local variables. Table 2.6 gives parameter estimates for global variables and Table 2.7 presents parameter estimates for local variables of the MGWR model. The AIC value of MGWR model is less than GWR model indicates that MGWR model performs better than GWR model.

Similarly, Figure 2.6 describes map of MGWR coefficients of local variables.

Table 2. 6 Coefficients of MGWR model (Global variables).

Variables	Estimates
Poverty	-0.02
Unemployment	0.07
German ancestry	0.003

Table 2. 7 Coefficients of MGWR model (Local variables).

Variables	Minimum	First Quartile	Median	Third Quartile	Maximum
Intercepts	1.18	5.44	7.13	9.92	16.5
Obesity	0.05	0.11	0.15	0.17	0.22
Physical inactivity	0.02	0.1	0.13	0.16	0.22
Nonwhite	-0.04	-0.007	0.002	0.03	0.08
Education	-0.07	-0.02	0.02	0.03	0.08
Labor force	-0.16	-0.11	-0.09	-0.06	-0.006
AICc	3042				

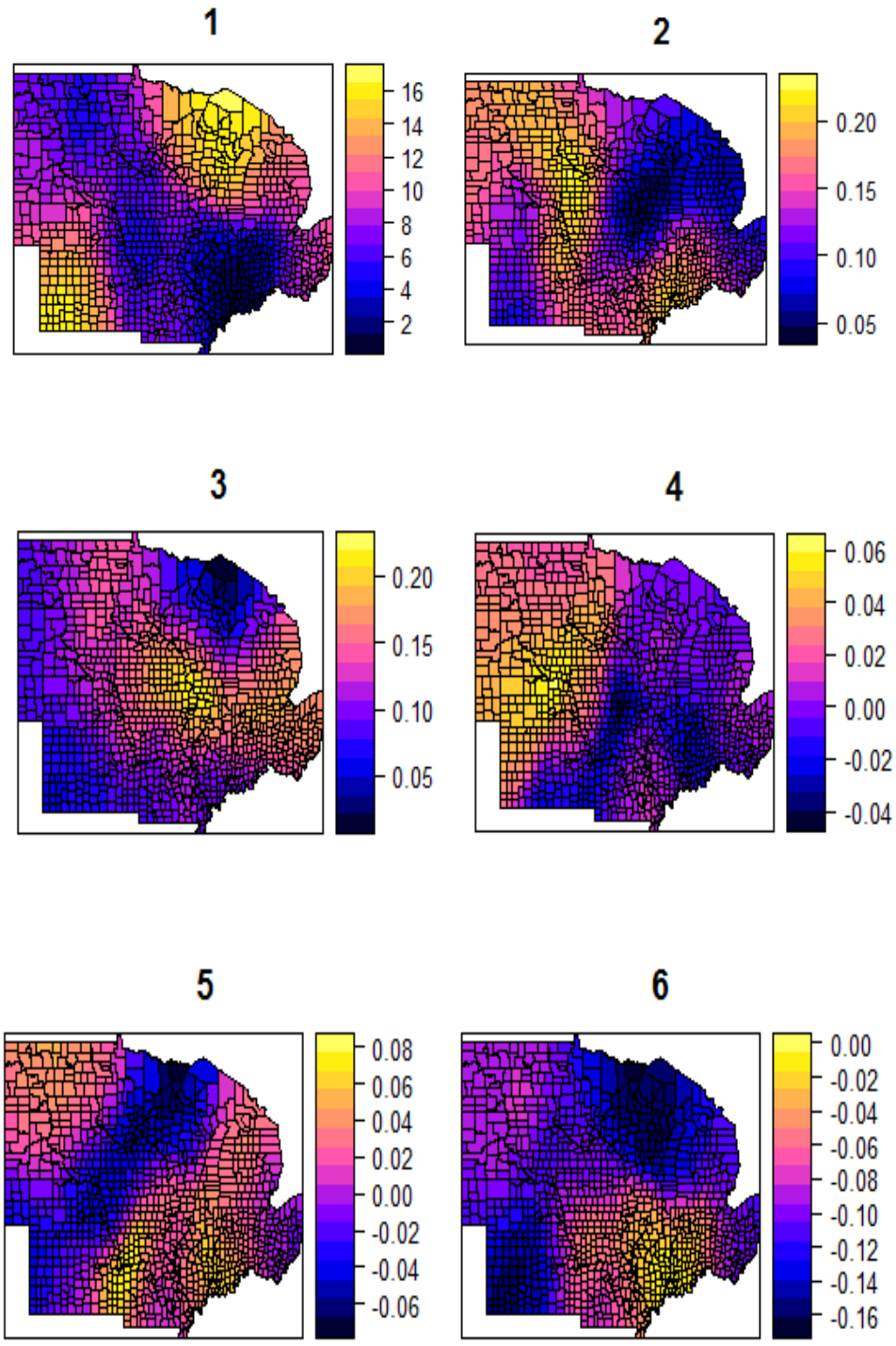


Figure 2. 6 MGWR coefficients of local variables: (1) Intercepts (2) Obesity (3) Physical inactivity (4) Nonwhites (5) Education (6) Labor force.

2.4 Discussion

Two basic assumptions of OLS model are the residuals are independent and have constant variance. The violation of these basic assumptions in spatial data will lead to erroneous results if we fit OLS model. Fotheringham et al. (1996) introduced GWR model to address the issue of spatial autocorrelation by fitting spatially varying local models. In the GWR model, variables are treated as spatially nonstationary that means there is a spatially varying relationship between dependent and independent variables. Sometimes the effect of some of the variables varies negligibly with geographical location while others vary geographically. Fotheringham et al. (2003) purposed to extend the GWR model to MGWR to address this issue by having both local and global variables into a single model.

The results of this study showed some of the variables are associated globally and others have local effects in diabetes prevalence. In recent years, there has been increasing research interest in effects of different variables on the geographical distribution of diseases. GWR model has been a famous choice of researchers who want to fit regression models with spatial data, but the result of this study showed that the effect of all the explanatory variables on diabetes prevalence does not vary spatially. Fitting GWR model with data with some of the variables with global effect will give erroneous results (Fotheringham et al., 2003).

The spatial distribution of diabetes prevalence and other independent variables varied strongly in this county-level data. This was reflected by the result of local Moran's I plots and MGWR. The issue of spatial heterogeneity is very common for other chronic diseases such as cancer (Fu, Jiang, Lin, Liu, & Wang, 2015; Goovaerts et al., 2015; Jia, James, &

Kedia, 2014; Ren et al., 2016; Yao, Foltz, Odisho, & Wheeler, 2015); heart diseases (Ford & Highfield, 2016; Hu, 2009; Lim et al., 2014; Srinivasan), obesity (Chalkias et al., 2013; Fraser et al., 2012; Procter, Clarke, Ransley, & Cade, 2008). Our results, therefore, supported the previous findings that chronic diseases are usually clustered spatially.

The results of this study show that there is a positive relationship between diabetes prevalence and obesity and the effect varies with the locations which is consistent with previous studies (Dijkstra et al., 2013; Hipp & Chalise, 2015) that fit GWR model on study of diabetes prevalence and obesity.

This study showed that physical inactivity was positively associated with diabetes prevalence rate and the effect varies spatially which is consistent with the similar studies (Hipp & Chalise, 2015) in the continental Unites States. Similarly, there was geographically varying association between nonwhites and diabetes prevalence. The association between diabetes prevalence and education varies geographically this finding is consistent with similar studies (Hipp & Chalise, 2015) in the continental United States. The effect of percent of people in labor force on diabetes prevalence also varied geographically. To the best of our knowledge, this is the first study that tested the relationship of diabetes prevalence with percentage of people in labor force.

The results of this study also showed that the effects of poverty, unemployment, and German ancestry on diabetes prevalence do not vary geographically. Previous study by (Hipp & Chalise, 2015) showed that the effect of poverty and unemployment varies with geographical locations in the continental United States. This result might be different because it includes data from counties in the Midwestern United States only. From a methodological aspect, MGWR proves to be the best model based upon AICc values of the

models. The main strength of this study is to use the MGWR to study the association between diabetes and socioeconomic and lifestyle factors. MGWR model is superior compared to OLS model and basic GWR model. The benefit of fitting geographical location specific models is it provides the local insight of the problems and helps public health policymakers to make effective policies to control and prevent diabetes.

2.5 Conclusion

Eight different risk factors of diabetes were identified in the Midwestern United States: (1) Obesity (2) Physical inactivity (3) Unemployment (4) Education (5) Poverty (6) Nonwhite (7) Labor force, and (8) German ancestry. The result supports the use of MGWR model better describes the relationship between diabetes prevalence and lifestyle and socioeconomic variables. The effect of poverty, unemployment, and German ancestry was global and the effects of the rest of the covariates on diabetes prevalence vary geographically. The use of MGWR can be useful to study spatial pattern of different diseases. The findings of this study can also be useful for policy makers to make effective policy based on geographical locations. Different strategies for diabetes reduction may be appropriate in different locations because of the spatially varying effects of covariates on diabetes prevalence.

References

Arslanian, S. A. (2000). Type 2 diabetes mellitus in children: pathophysiology and risk factors. *Journal of Pediatric endocrinology and Metabolism, 13*(Supplement), 1385-1394.

Arslanian, S. A., Bacha, F., Saad, R., & Gungor, N. (2005). Family history of type 2 diabetes is associated with decreased insulin sensitivity and an impaired balance between insulin sensitivity and insulin secretion in white youth. *Diabetes care, 28*(1), 115-119.

Benson, T., Chamberlin, J., & Rhinehart, I. (2005). An investigation of the spatial determinants of the local prevalence of poverty in rural Malawi. *Food Policy, 30*(5), 532-550.

Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1996). Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical Analysis, 28*(4), 281-298.

Brunsdon, C., Fotheringham, S., & Charlton, M. (2000). Geographically weighted regression as a statistical model.

Cahill, M., & Mulligan, G. (2007). Using geographically weighted regression to explore local crime patterns. *Social Science Computer Review, 25*(2), 174-193.

Chalkias, C., Papadopoulos, A. G., Kalogeropoulos, K., Tambalis, K., Psarra, G., & Sidossis, L. (2013). Geographical heterogeneity of the relationship between childhood obesity and socio-environmental status: Empirical evidence from Athens, Greece. *Applied Geography, 37*, 34-43.

Chen, V. Y.-J., Wu, P.-C., Yang, T.-C., & Su, H.-J. (2010). Examining non-stationary effects of social determinants on cardiovascular mortality after cold surges in Taiwan. *Science of the total environment, 408*(9), 2042-2049.

Chi, S.-H., Grigsby-Toussaint, D. S., Bradford, N., & Choi, J. (2013). Can geographically weighted regression improve our contextual understanding of obesity in the US? Findings from the USDA Food Atlas. *Applied Geography, 44*, 134-142.

Deller, S. (2010). Rural poverty, tourism and spatial heterogeneity. *Annals of Tourism Research, 37*(1), 180-205.

Dijkstra, A., Janssen, F., De Bakker, M., Bos, J., Lub, R., Van Wissen, L. J., & Hak, E. (2013). Using spatial analysis to predict health care use at the local level: a case study of type 2 diabetes medication use and its association with demograpHic change and socioeconomic status. *PLoS One, 8*(8), e72730.

Farrow, A., Larrea, C., Hyman, G., & Lema, G. (2005). Exploring the spatial variation of food poverty in Ecuador. *Food Policy, 30*(5), 510-531.

Ford, M. M., & Highfield, L. D. (2016). Exploring the Spatial Association between Social Deprivation and Cardiovascular Disease Mortality at the Neighborhood Level. *PLoS One, 11*(1), e0146085. doi:10.1371/journal.pone.0146085

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2003). *Geographically weighted regression: the analysis of spatially varying relationships*: John Wiley & Sons.

Fraser, L. K., Clarke, G. P., Cade, J. E., & Edwards, K. L. (2012). Fast food and obesity: a spatial analysis in a large United Kingdom population of children aged 13–15. *American journal of preventive medicine, 42*(5), e77-e85.

Fu, J., Jiang, D., Lin, G., Liu, K., & Wang, Q. (2015). An ecological analysis of PM $_{2.5}$ concentrations and lung cancer mortality rates in China. *BMJ Open, 5*(11). doi:10.1136/bmjopen-2015-009452

Gollini, I., Lu, B., Charlton, M., Brunsdon, C., & Harris, P. (2013). GWmodel: an R package for exploring spatial heterogeneity using geographically weighted models. *arXiv preprint arXiv:1306.0413*

Goovaerts, P., Xiao, H., Adunlin, G., Ali, A., Tan, F., Gwede, C. K., & Huang, Y. (2015). GEOGRAPHICALLY-WEIGHTED REGRESSION ANALYSIS OF PERCENTAGE OF LATE-STAGE PROSTATE CANCER DIAGNOSIS IN FLORIDA. *Appl Geogr, 62*, 191-200. doi:10.1016/j.apgeog.2015.04.018

Haire-Joshu, D. L. (2015). Next Steps: Eliminating Disparities in Diabetes and Obesity. *Preventing chronic disease, 12*.

Harris, M. I. (2001). Racial and ethnic differences in health care access and health outcomes for adults with type 2 diabetes. *Diabetes care, 24*(3), 454-459.

Hipp, J. A., & Chalise, N. (2015). Spatial Analysis and Correlates of County-Level Diabetes Prevalence, 2009–2010. *Preventing chronic disease, 12*.

Hu, Z. (2009). Spatial analysis of MODIS aerosol optical depth, PM2.5, and chronic coronary heart disease. *International Journal of Health Geographics, 8*(1), 27. doi:10.1186/1476-072x-8-27

Irigoien, X., Klevjer, T. A., Røstad, A., Martinez, U., Boyra, G., Acuña, J., . . . Hernandez-Leon, S. (2014). Large mesopelagic fishes biomass and trophic efficiency in the open ocean. *Nature communications, 5*.

Jia, C., James, W., & Kedia, S. (2014). Relationship of Racial Composition and Cancer Risks from Air Toxics Exposure in Memphis, Tennessee, U.S.A. *International journal of environmental research and public health, 11*(8), 7713-7724. doi:10.3390/ijerph110807713

Kang, D., & Dall'erba, S. (2016). Exploring the spatially varying innovation capacity of the US counties in the framework of Griliches' knowledge production function: a mixed GWR approach. *Journal of Geographical Systems, 18*(2), 125-157.

Klein, R. (1995). Hyperglycemia and microvascular and macrovascular disease in diabetes. *Diabetes care, 18*(2), 258-268.

Kuusisto, J., Mykkänen, L., Pyörälä, K., & Laakso, M. (1994). NIDDM and its metabolic control predict coronary heart disease in elderly subjects. *Diabetes, 43*(8), 960-967.

Lehto, S., Rönnemaa, T., Pyörälä, K., & Laakso, M. (1996). Predictors of stroke in middle-aged patients with non–insulin-dependent diabetes. *Stroke, 27*(1), 63-68.

Lim, Y.-R., Bae, H.-J., Lim, Y.-H., Yu, S., Kim, G.-B., & Cho, Y.-S. (2014). Spatial analysis of PM10 and cardiovascular mortality in the Seoul metropolitan area. *Environmental health and toxicology, 29*.

Lin, C.-H., & Wen, T.-H. (2011). Using geographically weighted regression (GWR) to explore spatial varying relationships of immature mosquitoes and human densities with the incidence of dengue. *International journal of environmental research and public health, 8*(7), 2798-2815.

Longley, P. A., & Tobón, C. (2004). Spatial dependence and heterogeneity in patterns of hardship: an intra-urban analysis. *Annals of the Association of American Geographers, 94*(3), 503-519.

Lu, B., Harris, P., Charlton, M., & Brunsdon, C. (2015). Calibrating a Geographically Weighted Regression Model with Parameter-specific Distance Metrics. *Procedia Environmental Sciences, 26*, 110-115.

Martínez-Fernández, J., Chuvieco, E., & Koutsias, N. (2013). Modelling long-term fire occurrence factors in Spain by accounting for local variations with geographically weighted regression. *Natural Hazards and Earth System Sciences, 13*(2), 311-327.

Mei, C. L., He, S. Y., & Fang, K. T. (2004). A Note on the Mixed Geographically Weighted Regression Model*. *Journal of Regional Science, 44*(1), 143-157.

Mitchell, M., & Yuan, F. (2010). Assessing forest fire and vegetation recovery in the Black Hills, South Dakota. *Giscience & Remote Sensing, 47*(2), 276-299.

Nakaya, T., Fotheringham, A., Brunsdon, C., & Charlton, M. (2005). Geographically weighted Poisson regression for disease association mapping. *Statistics in Medicine, 24*(17), 2695-2717.

Pirart, J. (1978). Diabetes mellitus and its degenerative complications: a prospective study of 4,400 patients observed between 1947 and 1973. *Diabetes care, 1*(3), 168-188.

Plant, R. E. (2012). *Spatial data analysis in ecology and agriculture using R*: cRc Press.

Procter, K., Clarke, G., Ransley, J., & Cade, J. (2008). Micro-level analysis of childhood obesity, diet, physical activity, residential socioeconomic and social capital variables: where are the obesogenic environments in Leeds? *Area, 40*(3), 323-340.

Ren, H., Cao, W., Chen, G., Yang, J., Liu, L., Wan, X., & Yang, G. (2016). Lung Cancer Mortality and Topography: A Xuanwei Case Study. *Int J Environ Res Public Health, 13*(5). doi:10.3390/ijerph13050473

Sá, A. C., Pereira, J. M., Charlton, M. E., Mota, B., Barbosa, P. M., & Fotheringham, A. S. (2011). The pyrogeography of sub-Saharan Africa: a study of the spatial non-stationarity of fire–environment relationships using GWR. *Journal of Geographical Systems, 13*(3), 227-248.

Schuurman, N., Peters, P. A., & Oliver, L. N. (2009). Are obesity and physical activity clustered? A spatial analysis linked to residential density. *Obesity, 17*(12), 2202-2209.

Selby, B., & Kockelman, K. M. (2013). Spatial prediction of traffic levels in unmeasured locations: applications of universal kriging and geographically weighted regression. *Journal of Transport Geography, 29*, 24-32.

Sheehan, K. R., Strager, M. P., & Welsh, S. A. (2013). Advantages of geographically weighted regression for modeling benthic substrate in two greater yellowstone ecosystem streams. *Environmental Modeling & Assessment, 18*(2), 209-219.

Siordia, C., Saenz, J., & Tom, S. E. (2012). An introduction to macro-level spatial nonstationarity: a geographically weighted regression analysis of diabetes and poverty. *Human geographies, 6*(2), 5.

Troy, A., Grove, J. M., & O'Neil-Dunne, J. (2012). The relationship between tree canopy and crime rates across an urban–rural gradient in the greater Baltimore region. *Landscape and Urban Planning, 106*(3), 262-270.

Tseng, C.-T., Su, N.-J., Sun, C.-L., Punt, A. E., Yeh, S.-Z., Liu, D.-C., & Su, W.-C. (2013). Spatial and temporal variability of the Pacific saury (Cololabis saira) distribution in the northwestern Pacific Ocean. *ICES Journal of Marine Science: Journal du Conseil, 70*(5), 991-999.

Turner, R., Millns, H., Neil, H., Stratton, I., Manley, S., Matthews, D., & Holman, R. (1998). Risk factors for coronary artery disease in non-insulin dependent diabetes mellitus: United Kingdom Prospective Diabetes Study (UKPDS: 23). *Bmj, 316*(7134), 823-828.

Wei, C.-H., & Qi, F. (2012). On the estimation and testing of mixed geographically weighted regression models. *Economic Modelling, 29*(6), 2615-2620.

Wen, T.-H., Chen, D.-R., & Tsai, M.-J. (2010). Identifying geographical variations in poverty-obesity relationships: empirical evidence from Taiwan. *Geospatial health, 4*(2), 257-265.

Wheeler, D. C., & Waller, L. A. (2009). Comparing spatially varying coefficient models: a case study examining violent crime rates and their relationships to alcohol outlets and illegal drug arrests. *Journal of Geographical Systems, 11*(1), 1-22.

Windle, M. J., Rose, G. A., Devillers, R., & Fortin, M.-J. (2009). Exploring spatial non-stationarity of fisheries survey data using geographically weighted regression (GWR): an example from the Northwest Atlantic. *ICES Journal of Marine Science: Journal du Conseil*, fsp224.

Yan, J., Shu, X., & Yuan, H. (2010). Relationship between spatial distribution of thief crime and geographical factors. *Journal of Tsinghua University (Science and Technology), 2*, 003.

Yang, T.-C., Wu, P.-C., Chen, V. Y.-J., & Su, H.-J. (2009). Cold surge: a sudden and spatially varying threat to health? *Science of the total environment, 407*(10), 3421-3424.

Yao, N., Foltz, S. M., Odisho, A. Y., & Wheeler, D. C. (2015). Geographic Analysis of Urologist Density and Prostate Cancer Mortality in the United States. *PLoS One, 10*(6), e0131578. doi:10.1371/journal.pone.0131578

Yu, D., Wei, Y. D., & Wu, C. (2007). Modeling spatial dimensions of housing prices in Milwaukee, WI. *Environment and Planning B: Planning and Design, 34*(6), 1085-1102.

Zhang, H., & Song, W. (2014). Addressing issues of spatial spillover effects and non-stationarity in analysis of residential burglary crime. *GeoJournal, 79*(1), 89-102.

Zhang, H., Zhang, J., Lu, S., Cheng, S., & Zhang, J. (2011). Modeling hotel room price with geographically weighted regression. *International Journal of Hospitality Management, 30*(4), 1036-1043.

Zhao, F., & Park, N. (2004). Using geographically weighted regression models to estimate annual average daily traffic. *Transportation Research Record: Journal of the Transportation Research Board*(1879), 99-107.

Chapter 3

3 Bayesian Spatiotemporal Zero-Inflated Models for Areal Count Data

ABSTRACT

Lyme disease is one of the most common vector born disease in the United States. Minnesota is one of the states in the United States that has the higher incidence of Lyme disease cases each year. Researchers are trying to find the relationship between Lyme disease and socioeconomic, climatic, landscape, and physical covariates. Even though Lyme disease is very common in Minnesota, it is not distributed evenly. There are many counties that have no reported cases of Lyme disease. The study of Lyme disease count involves excess zero counts and spatiotemporal correlation. Bayesian hierarchical models which rely on computationally expensive MCMC have been used extensively to address the presence of spatiotemporal correlation. Recently developed computationally efficient INLA approach is increasingly popular as an effective alternative of MCMC. In this chapter, the effect of climatic covariates on Lyme disease count in Minnesota was studied by using INLA approach and fond the best model which was identified based on DIC. The findings of this study will help to prevent and control Lyme disease.

3.1 Introduction

Vectors such as mosquitos, ticks, and fleas transmit vector borne diseases. In the United States, 14 different vector-borne diseases are of national public health concern. Climatic factors have a great impact on the seasonality, distribution, and prevalence of vector borne diseases such as Lyme disease (Gage, Burkot, Eisen, & Hayes, 2008). The geographic distribution of Lyme disease is limited to some specific areas in the United States, and there

is year-to-year variation in case count (Mead, 2015; Moore, Eisen, Monaghan, & Mead, 2014).

The climate patterns have great influence on survival and distribution primary host of Lyme disease *Ixodes scupularis* (Brownstein, Holford, & Fish, 2005; Johnson et al., 2016; Lindsay et al., 1995; Stafford, 1994). The variation in temperature and precipitation play important role in distribution of Lyme disease (Eisen, Eisen, & Beard, 2016; McCabe & Bunnell, 2004; N. Ogden et al., 2004; N. H. Ogden et al., 2008).

Bayesian approach is very effective in spatiotemporal data analysis in which we need to consider the spatial and temporal structure of data in the inferential process (Blangiardo, Cameletti, Baio, & Rue, 2013). Bayesian approach has been applied in several epidemiological applications (Bernardinelli et al., 1995; Best, Richardson, & Thomson, 2005; Lawson, 2013; Musenge, Chirwa, Kahn, & Vounatsou, 2013). For example, we can specify disease mapping and/or ecological regression if the data is aggregated counts of outcomes and covariates, alternatively, we can use geostatistical models if we data are observed at point locations (Blangiardo et al., 2013).

Hierarchical Bayesian models rely on computationally expensive and technically challenging MCMC simulation techniques. A novel methodology, INLA, has been developed as an alternative of MCMC. INLA method use approximation techniques for inference that helps to avoid intense computational demands, convergence and mixing problems (Blangiardo et al.,2015).

It is common to have a large proportion of data with zeros in ecological, epidemiological and environmental studies (Arab, 2015). There are varieties of ways to model count data

such as Poisson, negative binomial, binomial etc. Poisson regression is the most commonly used model for spatial count data (Agarwal, Gelfand, & Citron-Pousty, 2002).

The study of multiple climatic factors that contribute to the increased risk of Lyme disease helps to control and spread of the disease. Since the forecast for average climatic conditions for forthcoming weeks, months and seasons is available, the finding of this study will help in Lyme disease control and prevention effort.

The objective of this study was to review the existing methodology of zero-inflated spatiotemporal data and find the spatiotemporal relationship between Lyme disease count data and climatic covariates in Minnesota.

3.2 Methods

3.2.1 Data

This study modeled Lyme disease count data in Minnesota. Data of confirmed cases of Lyme disease from 2008 to 2014 at the county-level for the state of Minnesota was obtained from the Center for Disease Control and Prevention (CDC). County-level temperature and precipitation data are obtained from PRISM weather data (htpp://prism.oregonstate.edu). County wise surveillance of Lyme disease in the United States is publicly available from CDC for the year 2000 to 2014. The national surveillance of Lyme disease was changed in 2008. Before 2008, Lyme disease case is confirmed by the presence of erythemia migrans (EM) rash or presence of one late stage symptoms with laboratory confirmation. However, after 2008, known exposure or the presence of EM rash only is sufficient to declare confirmed case of Lyme disease when it occurred in an already endemic county. A positive laboratory test is required for non-endemic counties (Li et al., 2014). This study only

included data of Lyme disease count in the state of Minnesota from 2008-2014 due to change in national surveillance of Lyme disease in 2008.

3.2.2 Spatiotemporal data

Spatiotemporal data have information about spatial region and time domain. Consider a spatial domain $\boldsymbol{S}$= {1, …, N} where i = 1, …, N are the index of areas for N number of areas. The neighbors of area i are denoted by N_i, for $i \in \boldsymbol{S}$. That is,

$N_i = \{j \in \boldsymbol{S}: j \text{ is neighbor of } i\}, i \in \boldsymbol{S}$.

The neighbors of a given area can be defined by adjacency criteria such as sharing common border. The number of neighbors based on adjacency criterion for the i^{th} area is denoted as n_i. Total T time points are index by t as 1, 2, 3, … , T. The response variable y_{it} denotes count of area i at time t, where i = 1,2,3, … , N, and t = 1,2,3, …. , T. The values for covariate k for area i, at time point t is denoted by x_{itk} where i = 1,2,3, … , N, t = 1,2,3, …, T, and k = 1,2,3, …., K. The covariates can be written as a vector form $\mathbf{x_{it}}$, that is: $\mathbf{X_{it}} = (x_{it1}, x_{it2}, x_{it3}, ..., x_{itk})'$.

3.2.3 Bayes' Theorem

Bayes' theorem states that:

$$P(\theta \mid y) = \frac{P(y \mid \theta) \times P(\theta)}{P(y)} \quad \ldots\ldots\ldots (1)$$

Where $P(\theta \mid y)$ is posterior probability density, $P(y \mid \theta)$ is likelihood, $P(\theta)$ is prior and $P(y)$ is normalizing constant.

Since the denominator does not contain θ, it acts as normalizing constant. Therefore, we can simplify the Bayes Theorem as:

$$\text{Posterior} \propto \text{Likelihood x Prior}$$

3.2.4 Hierarchical model

The basic hierarchical model consists of three primary stages: (1) Data model ([data|process, parameters]); (2) process model ([process|parameters]); (3) parameter model where the brackets and vertical line refers to probability distribution and conditional distribution respectively. This approach is helpful in data analysis because it breaks the complex statistical modeling problem into pieces (Wikle & Anderson, 2003).

The first stage of the hierarchical model involved with data observational process, or data model, which represents the distribution of the data (example; counts of Lyme disease) given the process of interest (example; count of Lyme disease over some geographic area) and the parameters that describe the data model. The second stage of the model describes the process conditional on other parameters. For example, this might be a regression model relating Lyme disease count data to some climatic covariates with model parameters represents the strength of the association. The last stage accounts for uncertainty in parameters by assigning them distributions. For example, the effect of climatic covariates could be different at different geographical location and time. Thus, we might allow the regression parameters varying spatially and temporally by assigning distributions, that includes spatiotemporal correlation.

The benefit of fitting hierarchical model is it allows parameters to vary more than one level via the introduction of random effect that helps to simplify the complex interactions.

Application of hierarchical Bayesian models is very useful for modeling complex data structure for example explicit modeling of spatiotemporal variability (Cressie, Calder, Clark, Hoef, & Wikle, 2009).

3.2.5 Zero-inflated count data models

A variety of modeling approaches are available to model spatial and spatiotemporal count data. For example; count data are modeled by using Poisson, negative binomial, binomial, beta binomial or hyper-geometric distributions (Agarwal et al., 2002; Lambert, 1992). Poisson regression is the most frequently used method for the spatial count data.

$y_i \sim$ Poisson (λ_i) Where i= 1,…, n.

Then, E (y_i) = variance (y_i) = λ_i.

We can show this relationship upon considering the probability mass function

$$f(y_i;\lambda_i)=\frac{e^{-\lambda_i}\lambda^{y_i}}{y_i!}, \text{ for i = 1,… , n.}$$

From this, we can show that

$$E(y_i)=\sum_{y_i=0}^{\infty} y_i \frac{\exp(-\lambda_i)\lambda_i^{\,y_i}}{y_i!}=\sum_{y_i=1}^{\infty} y_i \frac{\exp(-\lambda_i)\lambda_i^{\,y_i}}{y_i!}$$

$$=\sum_{y_i=1}^{\infty}\frac{\exp(-\lambda_i)\lambda_i^{\,y_i-1}\lambda_i}{(y_i-1)!}=\exp(-\lambda_i)\lambda_i\sum_{y_i=1}^{\infty}\frac{\lambda_i^{\,y_i-1}}{(y_i-1)!}$$

$$=\exp(-\lambda_i)\lambda_i\left(\frac{\lambda_i^0}{0!}+\frac{\lambda_i^1}{1!}+\frac{\lambda_i^2}{2!},...\right)=\exp(-\lambda_i)\lambda_i\sum_{y_{i=0}}^{\infty}\frac{\lambda_i^{y_i}}{y_i!}$$

$$= \lambda_i \sum_{y_{i=0}}^{\infty} \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!} = \lambda_i * 1$$

$$= \lambda_i$$

$$Var(y_i) = E(y_i)^2 - (E(y_i))^2$$

$$= E(y_i)^2 - (E(y_i))^2$$

$$= E((y_i)(y_i - 1) + y_i)) - (E(y_i))^2$$

$$E((y_i)(y_i - 1)) = \sum_{y_{i=0}}^{\infty} (y_i)(y_i - 1) \frac{\exp(-\lambda_i)\lambda^{y_i}}{y_i!}$$

$$= \sum_{y_{i=2}}^{\infty} \frac{\exp(-\lambda_i)\lambda^{y_i}}{(y_i - 2)!} = \lambda_i^2 \exp(-\lambda_i) \sum_{y_{i=2}}^{\infty} \frac{\lambda^{y_i - 2}}{(y_i - 2)!}$$

$$= \lambda_i^2 \exp(-\lambda_i) \sum_{y_{i=2}}^{\infty} \frac{\lambda^{y_i - 2}}{(y_i - 2)!} = \lambda_i^2 \exp(-\lambda_i) \left(\frac{\lambda^0}{0!} + \frac{\lambda^1}{1!} + ... \right)$$

$$= \lambda_i^2 \sum_{y_{i=0}}^{\infty} \frac{\exp(-\lambda_i)\lambda^{y_i}}{y_i!} = \lambda_i^2 * 1$$

$$= \lambda_i^2$$

$$Var = \lambda_i^2 + \lambda_i - \lambda_i^2$$

$$= \lambda_i$$

The most commonly used zero-inflated mixture models for count data are zero-inflated Poisson mixture model (ZIP) and zero-inflated negative binomial mixture model (ZINB) (Agarwal et al., 2002; Lambert, 1992).

3.2.6 Zero-inflated data models

Let Y denotes the observed variable. We assume the following model

$P^Y \in \{P_\theta, \theta = (p, \lambda)\}$ for Y.

$$Y = \Delta Z_1 + (1-\Delta)Z_2$$

Where Δ ~Bernoulli (p), $Z_1 \sim P^{Z_1} \in \{P_\lambda\}$, and $Z_2 \sim \delta_{\{0\}}$ are independent. We also assume that Z_1 is discrete and P ($Z_1 \geq 0$) =1.We can apply the above described model for count data because the observed variable Y has non-negative support.

$$P(Y = y) = \begin{cases} (1-p) + pp_{Z_1}(0), & y = 0 \\ pp_{Z_1}(y), & y = 1,2,3, \ldots \end{cases}$$

If Z ~ Poisson (λ) we obtain Zero-inflated Poisson model as

$$P(Y = y) = \begin{cases} (1-p) + pexp(-\lambda), & y = 0 \\ p\dfrac{\lambda^y \exp(-\lambda)}{y!}, & y = 1,2,3, \ldots \end{cases}$$

Theorem 1 and Theorem 2 describes various properties of zero-inflated data (Eggers, 2015).

Theorem 1. The excepted value E[Y] and variance Var [Y] of Y are given by E[Y] = $pE[Z_1]$ and Var[Y] = $pVar[Z_1] + (1-p)p\, E[Z_1]^2$.

Proof. The expected value of Y is given by

$$E[Y] = E[\Delta Z_1 + (1-\Delta)Z_2)]$$

$$= E[\Delta Z_1] + E[(1-\Delta)Z_2)]$$

$$= E[\Delta Z_1] + E[Z_2] - E[\Delta Z_2)]$$

$$= E[\Delta]E[Z_1] + E[Z_2] - E[\Delta]E[Z_2]$$

$$= E[\Delta]E[Z_1]$$

$$= pE[Z_1]$$

The variance of Y is given by

$$Var[Y] = Var[\Delta Z_1 + (1-\Delta)Z_2)]$$

$$= Var[\Delta Z_1] + Var[Z_2] + Var[\Delta Z_2]$$

$$= Var[\Delta Z_1] + Var[\Delta Z_2]$$

$$= E[\Delta^2]E[Z_1^2] - E[\Delta]^2 E[Z_1]^2 + E[\Delta^2]E[Z_2^2] - E[\Delta]^2 E[Z_2]^2$$

$$= E[\Delta^2]E[Z_1^2] - E[\Delta]^2 E[Z_1]^2$$

$$= (E[\Delta^2] - E[\Delta]^2 + E[\Delta]^2)(E[Z_1^2] - E[Z_1]^2 + E[Z_1]^2] - E[\Delta]^2 E[Z_1]^2$$

$$= (Var[\Delta] + E[\Delta]^2)(Var[Z_1] + E[Z_1]^2) - E[\Delta]^2 E[Z_1]^2$$

$$= (p(1-p) + p^2)(Var[Z_1] + E[Z_1]^2) - p^2 E[Z_1]^2$$

$$= (p(1-p) + p^2)(Var[Z_1] + E[Z_1]^2) - p^2 E[Z_1]^2$$

$$= pVar[Z_1] + pE[Z_1]^2 - p^2 E[Z_1]^2$$

$$= pVar[Z_1] + p(1-p)E[Z_1]^2$$

Corollary. The expected value E[Y] and variance Var[Y] of Poisson model are given by E[Y]= pλ, and var[Y]= pλ(1+λ-pλ).

Proof: In zero-inflated Poisson model, Z_1~ Poisson (λ), and $E(Z_1) = Var(Z_1) = \lambda$. Plugging these values for $E(Z_1)$ and $Var(Z_1)$ in the expression will give above result.

Theorem 2. Let Z_1 ~ Poisson (λ). The moment estimators $\hat{p}_{MME}(Y)$ and $\hat{\lambda}_{MME}$ are given by

$$\hat{p}_{MME}(Y) = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i}{\frac{\sum_{i=1}^{n} Y_i^2 - \sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} Y_i}} \text{ and}$$

$$\hat{\lambda}_{MME} = \frac{\sum_{i=1}^{n} Y_i^2}{\sum_{i=1}^{n} Y_i} - 1$$

The moment estimators of $\hat{p}_{MME}(Y)$ and $\hat{\lambda}_{MME}$ (Y) are given by the values of p and λ which satisfy

$$E[Y] = \frac{1}{n}\sum_{i=1}^{n} Y_i = p\lambda$$

$$E[Y^2] = \frac{1}{n}\sum_{i=1}^{n} Y_i^2$$

$$= Var[Y] + E[Y]^2$$

$$= p\lambda(1 + \lambda - p\lambda) + p^2\lambda^2$$

$$= p\lambda(1+\lambda)$$

$$\Rightarrow 1+\lambda = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i^2}{\frac{1}{n}\sum_{i=1}^{n} Y_i} = \frac{\sum_{i=1}^{n} Y_i^2}{\sum_{i=1}^{n} Y_i}$$

$$\Rightarrow \lambda = \frac{\sum_{i=1}^{n} Y_i^2}{\sum_{i=1}^{n} Y_i} - 1$$

$$\Rightarrow p = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i}{\lambda} = \quad = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i}{\frac{\sum_{i=1}^{n} Y_i^2}{\sum_{i=1}^{n} Y_i} - 1} = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i}{\frac{\sum_{i=1}^{n} Y_i^2 - \sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} Y_i}}$$

3.2.7 Latent Gaussian Model

The INLA framework deals with latent Gaussian model, where response variable $\mathbf{y_i}$ and the parameter of the family of distribution $\boldsymbol{\phi_i}$ is linked to a structured additive predictor $\boldsymbol{\eta_i}$ through link function **g (.)** so that $\mathbf{g}(\boldsymbol{\phi_i}) = \boldsymbol{\eta_i}$ (Martins, Simpson, Lindgren, & Rue, 2013). The INLA method is useful to estimate the effect of set of relevant covariates on some function (typically mean) of observed data with spatial or spatiotemporal correlation is taken into consideration in modeling.

The modelling framework for estimating the mean for the i^{th} unit by means of an additive predictor can be expressed as:

$$\eta_i = \alpha + \sum_{k=1}^{k} \beta_k x_{ki} + \sum_{l=1}^{L} f_l(z_{li}) \ldots\ldots\ldots (2)$$

Where,

α is the intercept, $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, ..., \beta_k)$ is the coefficients for the effects of some covariates $\mathbf{x} = (x_1, x_2, x_3, ..., x_k)$ on response, and **f** = {$f_1(.)$, …., $f_l(.)$} is a collection of functions defined in terms of a set of covariates **z** = (z_1, … , z_l). This formulation can accommodate a wide range of models from standard and linear hierarchical regression to spatial and spatiotemporal models by varying the form of function $\mathbf{f_l(.)}$ (Blangiardo et al., 2013).

The vector of parameters in (2) is represented by **θ = {α, β, f}.** We can assume GMRF prior on **θ,** with mean **0** and a precision matrix **Q** which reflects neighborhood structure. Q_{ij} = -1 if i and j are neighbors, and 0 otherwise. The diagonal elements of **Q** is $Q_{ii} = n_i$ where n_i represents the neighbors of the i^{th} area. The vector of K hyper-parameters $\boldsymbol{\psi} = (\psi_1, ..., \psi_K)$ which is much smaller than **θ**.

For example, if we consider 3 x 3 spatial grid as:

1	**2**	**3**
4	**5**	**6**
7	**8**	**9**

The Q matrix can be written as:

$$Q = \begin{pmatrix} 2 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 3 & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & 0 & 0 & -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 3 & -1 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & -1 & 4 & -1 & 0 & -1 & 0 \\ 0 & 0 & -1 & 0 & -1 & 3 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 0 & 0 & 2 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & -1 & 3 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 2 \end{pmatrix}$$

The temporal precision matrix for 7 time points represented by W as

$$W = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix}$$

The marginal posterior distributions for each of the parameters vector is computed by Bayesian computation as:

$$p(\theta_i \mid \mathbf{y}) = \int p(\boldsymbol{\psi} \mid \mathbf{y})(\theta_i \mid \boldsymbol{\psi}, \mathbf{y})\, d\boldsymbol{\psi}$$

and each element of the hyper-parameter vector as

$$p(\psi_k \mid \mathbf{y}) = \int p(\boldsymbol{\psi} \mid \mathbf{y})\, d\boldsymbol{\psi}_{-k}$$

We need to compute:

(1) $p(\boldsymbol{\psi} \mid \mathbf{y})$ to get all the marginals $p(\psi_k \mid \mathbf{y})$

(2) $p(\theta_i \mid \boldsymbol{\psi}, \mathbf{y})$ Which is needed to compute the marginal posterior for the parameters.

The first step of INLA method is to compute an approximation to the posterior marginal distribution of marginal hyper-parameters as

$$p(\boldsymbol{\psi} \mid \mathbf{y}) = \frac{p(\boldsymbol{\theta}, \boldsymbol{\psi} \mid \mathbf{y})}{p(\boldsymbol{\theta} \mid \boldsymbol{\psi}, \mathbf{y})} \quad \alpha \quad \frac{p(\boldsymbol{\psi})\, p(\boldsymbol{\theta} \mid \boldsymbol{\psi})\, p(\mathbf{y} \mid \boldsymbol{\theta})}{p(\boldsymbol{\theta} \mid \boldsymbol{\psi}, \mathbf{y})}$$

$$\approx \left. \frac{p(\boldsymbol{\psi})\, p(\boldsymbol{\theta} \mid \boldsymbol{\psi})\, p(\mathbf{y} \mid \boldsymbol{\theta})}{\tilde{p}(\boldsymbol{\theta} \mid \boldsymbol{\psi}, \mathbf{y})} \right|_{\theta = \theta^*(\psi)} =: \tilde{p}(\boldsymbol{\psi} \mid \mathbf{y}) \quad \ldots\ldots\ldots (3)$$

Where $\tilde{p}(\boldsymbol{\theta} \mid \boldsymbol{\psi}, \mathbf{y})$ is Gaussian approximation of $p(\boldsymbol{\theta} \mid \boldsymbol{\psi}, \mathbf{y})$ and $\boldsymbol{\theta}^*(\boldsymbol{\psi})$ is its mode.

The second step is slightly more complex because there are generally more elements in $\boldsymbol{\theta}$ than there are in $\boldsymbol{\psi}$ and this is more expensive in computation. An easy alternative to solve this issue is to approximate the posterior conditional distribution $p(\boldsymbol{\theta}_i \mid \boldsymbol{\psi}, \mathbf{y})$ directly as the marginal from, by using a Normal distribution where the precision matrix is based on the Cholesky decomposition of precision matrix $\mathbf{Q}$ (Rue & Martino, 2007). This method is very fast but less accurate. The alternative way is to re write the vector of parameters as $\boldsymbol{\theta} = (\theta_i, \boldsymbol{\theta}_{-i})$ and use Laplace approximation again to get

$$p(\boldsymbol{\theta}_i \mid \boldsymbol{\psi}, \mathbf{y}) = \frac{p(\theta_i, \boldsymbol{\theta}_{-i} \mid \boldsymbol{\psi}, \mathbf{y})}{p(\boldsymbol{\theta}_{-i} \mid \theta_i, \boldsymbol{\psi}, \mathbf{y})}$$

$$\approx \left. \frac{p(\boldsymbol{\theta}, \boldsymbol{\psi} \mid \mathbf{y})}{\tilde{p}(\boldsymbol{\theta}_{-i} \mid \theta_i, \boldsymbol{\psi}, \mathbf{y})} \right|_{\boldsymbol{\theta}_{-i} = \boldsymbol{\theta}_{-i}^*(\theta_i, \psi)} =: \tilde{p}(\boldsymbol{\theta}_i \mid \boldsymbol{\psi}, \mathbf{y}) \quad \ldots\ldots\ldots (4)$$

Random variable $(\boldsymbol{\theta}_{-i} \mid \theta_i, \boldsymbol{\psi}, \mathbf{y})$ are generally normally distributed and the approximation provided by (4) typically works well. This strategy also very expensive in terms of

computation. Better alternative is "Simplified Laplace Approximations" which is Taylor series expansion of the Laplace approximation. It requires shorter computation time and produces more accurate results.

INLA method starts with exploring the marginal joint posterior for the hyper-parameters $\tilde{p}(\boldsymbol{\psi} \mid \mathbf{y})$ to locate the mode; a grid search is then performed to get a set of relevant points $\{\psi_k\}$ along with a corresponding set of weights $\{\boldsymbol{\Delta}_K\}$, to give approximation to this distribution. For each ψ_k, the conditional posteriors $\tilde{p}(\theta_i \mid \psi_k, \mathbf{y})$ are obtained by numerical integration as:

$$\tilde{p}(\theta_i \mid \mathbf{y}) \approx \sum_{k=1}^{K} \tilde{p}(\theta_i \mid \psi_k, \mathbf{y})\, \tilde{p}(\psi_k \mid \mathbf{y})\, \boldsymbol{\Delta}_K$$

3.2.8 Count data modeling using INLA

County- level Lyme disease count data in Minnesota was spatially modeled by using INLA with Besag-Yourk-Mollie (BYM) method as described by Lawson (2013) as:

$$y_i = Poisson(\lambda_i = e_i \theta_i)$$

$$\log(\theta_i) = \boldsymbol{\beta} x_i + u_i + v_i$$

$$u \sim N(0, \tau_u)$$

$$v \sim N(\bar{v}_\delta, \frac{\tau_v}{v_\delta})$$

Where,

(1) y_i is the count of Lyme disease in county i; e_i is the expected count in county i; and θ_i is the relative risk of county i.

$$e_i = \left(\frac{\sum_i y_i}{\sum_i y_i^{\,p}} \right) \times y^p$$

y_i^p is the population in the i^{th} county of Minnesota.

$$\widehat{\theta_i} = \frac{y_i}{e_i}$$

(2) u_i is the spatially unstructured random effect component normally distributed with mean zero.

(3) vi is the spatially structured component which is modeled by using an intrinsic conditional autoregressive structure (iCAR) as:

$$v \sim N\left(\bar{v}_\delta, \frac{\tau_v}{v_\delta} \right)$$

Where, Neighborhood consist of spatially adjacent shapes is characterized by the normally distributed mean of the spatially structured random effect terms for the spatial shapes that makes the neighborhood $(\bar{v}_\delta)$ and the standard deviation of that mean divided by the number of the spatial shapes in the neighborhood $\left(\frac{\tau_v}{v_\delta} \right)$.

The spatial model described above can be extended to include temporal characteristics for a space-time model for count data in small areas in fixed time points. This approach extends Beasg-York-Mollie (BYM) by including a linear term for space time interaction and a nonparametric spatiotemporal time trend (Balngiardo et al., 2013). The spatiotemporal model can be written as:

$$\log(\theta_{ij}) = \beta_0 + \boldsymbol{\beta} x_i^T + u_i + v_i + \gamma_i + \omega_{ij}$$

Where β_0 is the intercept, $\boldsymbol{\beta} x_i^T$ is vector of coefficients for climatic covariates, u_i is spatially unstructured random effect term, v_i is the spatially structured conditional autoregressive term, γ_i is the first-order random walk-correlated time variable, and ω_{ij} is the interaction term for space and time for i= 1 … N small areas (N=87 counties in Minnesota) j= 1…T time points (T=7 years).

Spatiotemporal modeling was conducted using INLA approach with R-INLA package.

3.2.9 Bayesian Model selection using the Deviance Information Criterion (DIC)

Several models can be considered for a given data analysis. A large model fits data better because it has more flexibility, but larger models are difficult to compute and interpret. Choosing better model from competing models is also a very important issue in data analysis.

Some of the commonly used model selection criterion are Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and the Deviance Information Criterion (DIC).

Akaike Information Criterion (AIC) takes the form:

$$AIC = -2l_{mi}(\hat{\boldsymbol{\theta}}_{\mathbf{i}}) + 2p$$

Bayesian Information Criterion (BIC) takes the form:

$$BIC = -2l_{mi}(\hat{\boldsymbol{\theta}}_{\mathbf{i}}) + \log(n)p$$

Where $l_{m_i}(\hat{\boldsymbol{\theta}})$ is the log likelihood of the model m_i, $\hat{\boldsymbol{\theta}}_i$ is the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$ under the model m_i, n is the number of observations and p is the number of parameters. Models with lower AIC and BIC values are considered as better models. The number of parameters p in model determines the penalty for model complexity in both AIC and BIC calculation. AIC and BIC methods are not appropriate for the Bayesian hierarchical model where parameters include correlated random effects.

DIC, which is an extension of AIC, may be applied in choosing hierarchical spatial models. The criteria is defined as:

$$D(\theta) = -2\log L(\mathbf{y}|\boldsymbol{\theta})$$

Where $L(\mathbf{y}|\boldsymbol{\theta})$ is the likelihood function of the data given the parameters under the model. Then the DIC is defined as

$$DIC = \overline{D(\boldsymbol{\theta})} + p_D$$

Where $\overline{D(\boldsymbol{\theta})} = E[D(\theta)|\mathbf{y}]$ is the posterior mean of the deviance, a measure of fit with lower value indicates the better fit of the data. p_D is a penalty term which measures the complexity of the model and is defined as:

$$p_D = \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}})$$

Where $D(\bar{\boldsymbol{\theta}})$ is the deviance evaluated at the posterior mean of $\boldsymbol{\theta}$. p_D penalty accounts for spatial correlation or shrinkage among correlated parameters and gives estimate of effective number of model parameters rather than simply penalizing the models depends

on the total raw numbers parameters appearing in the model. The models with lower DIC score is preferred since it represents the best combination of fit and parsimony.

3.3 Results

Local Moran's I test was used to find local spatial autocorrelation. Figure 3.1 and Figure 3.2 display hot spots and cold spots of Lyme disease in Minnesota.

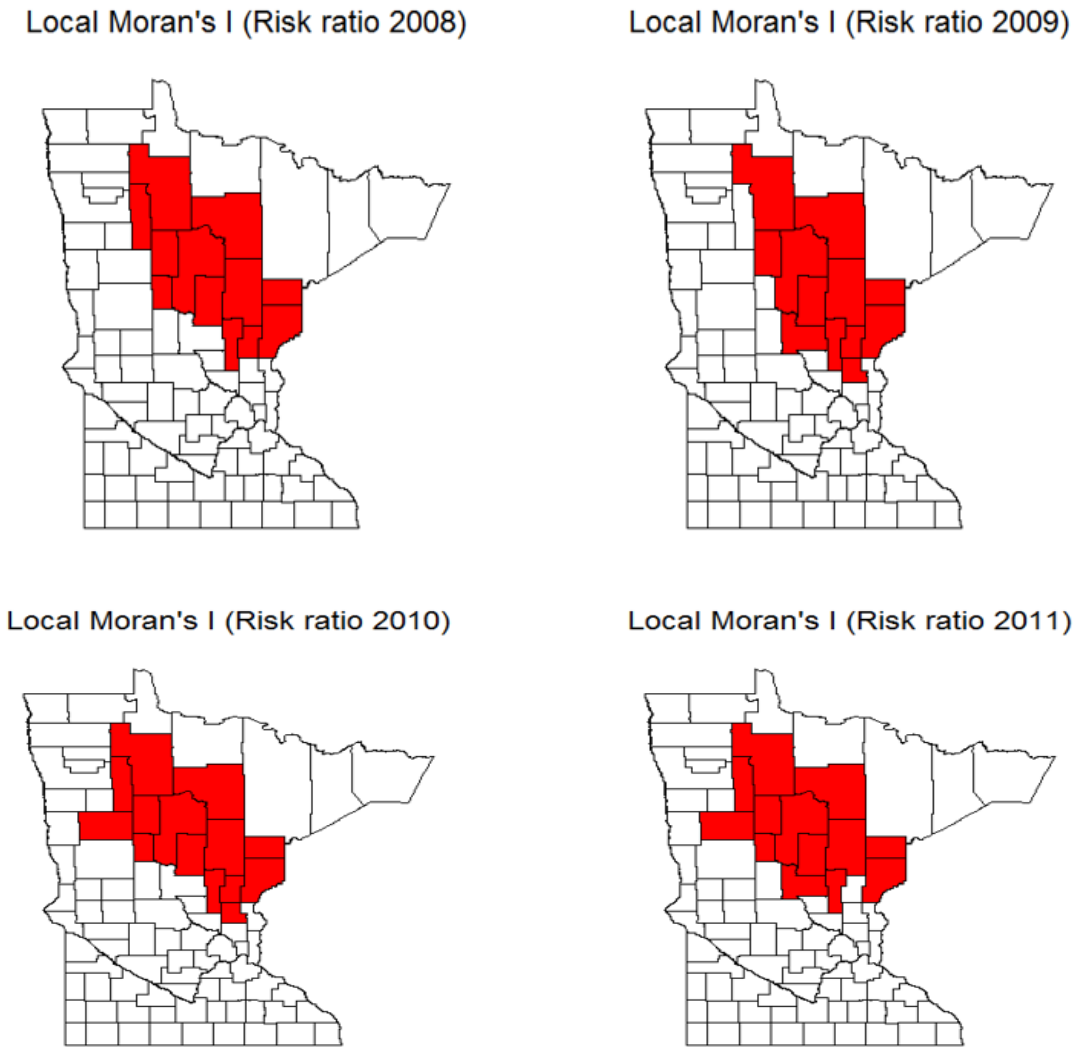


Figure 3. 1 Spatial cluster of Lyme disease in Minnesota for years 2008-2011.

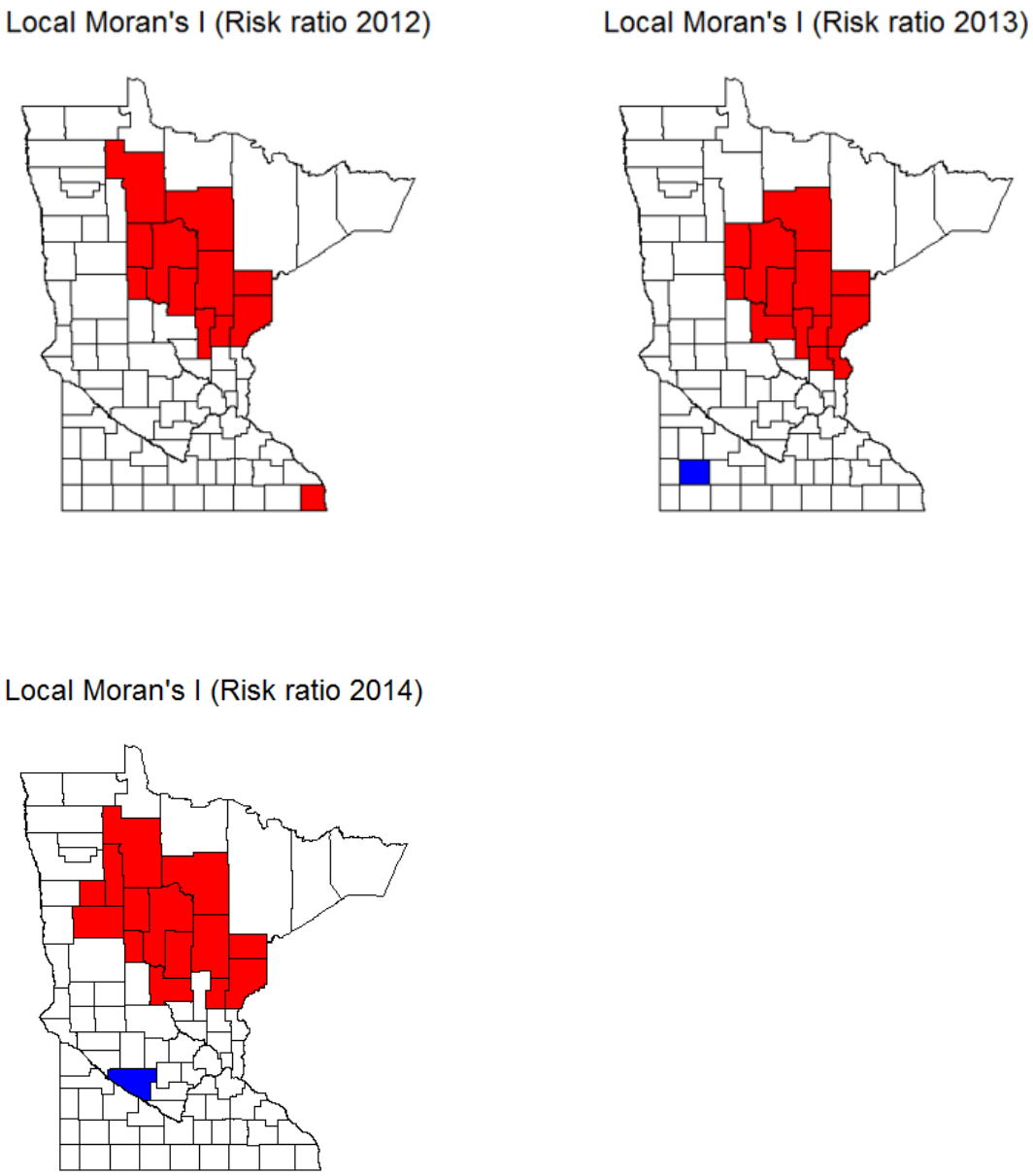


Figure 3. 2 Spatial cluster of Lyme disease in Minnesota for years 2012-2014.

Figure 3.3 and 3.4 display that a large number of counties in Minnesota have no recorded cases of Lyme disease.

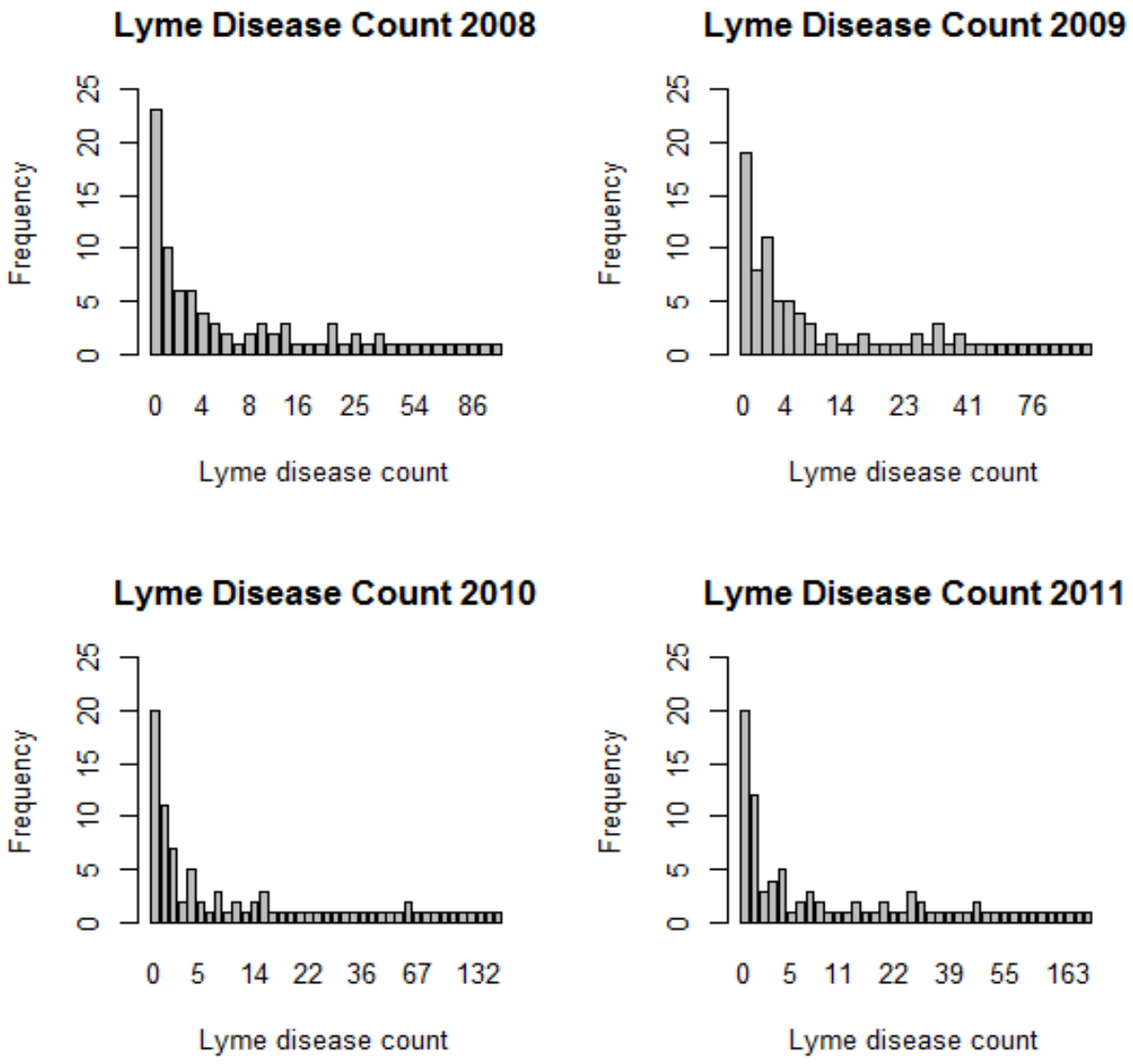


Figure 3. 3 Bar plot of Lyme disease count in Minnesota from year 2008-2011.

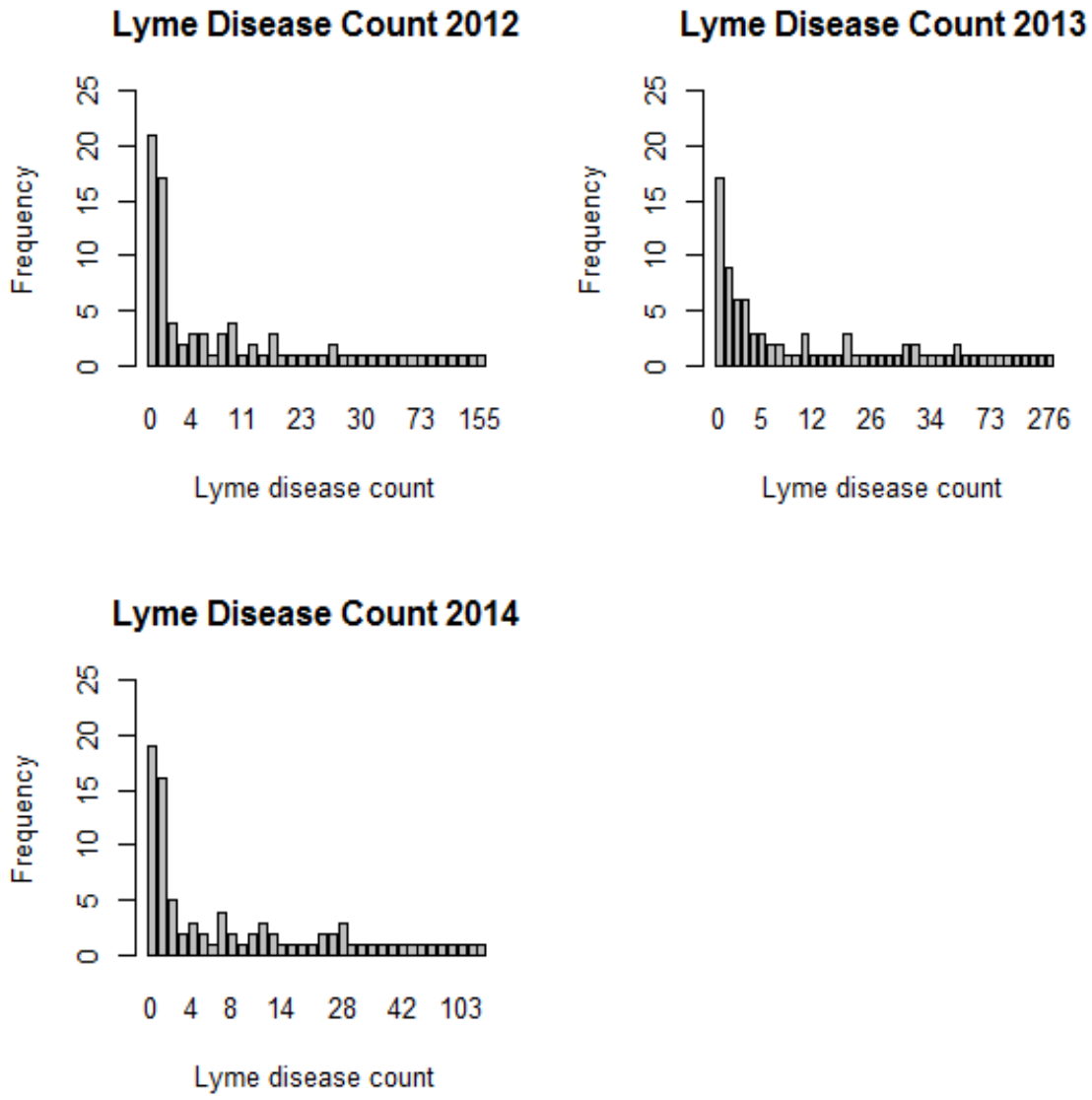


Figure 3. 4 Bar plot of Lyme disease count in Minnesota from year 2012-2014.

Figure 3.5 shows that there were fewer cases of Lyme disease in 2008 and 2014 compared to other years. There were many outliers in each year. Outliers are defined as any data value which is 1.5 interquartile rage (IQR) below the first quartile and above the third quartile.

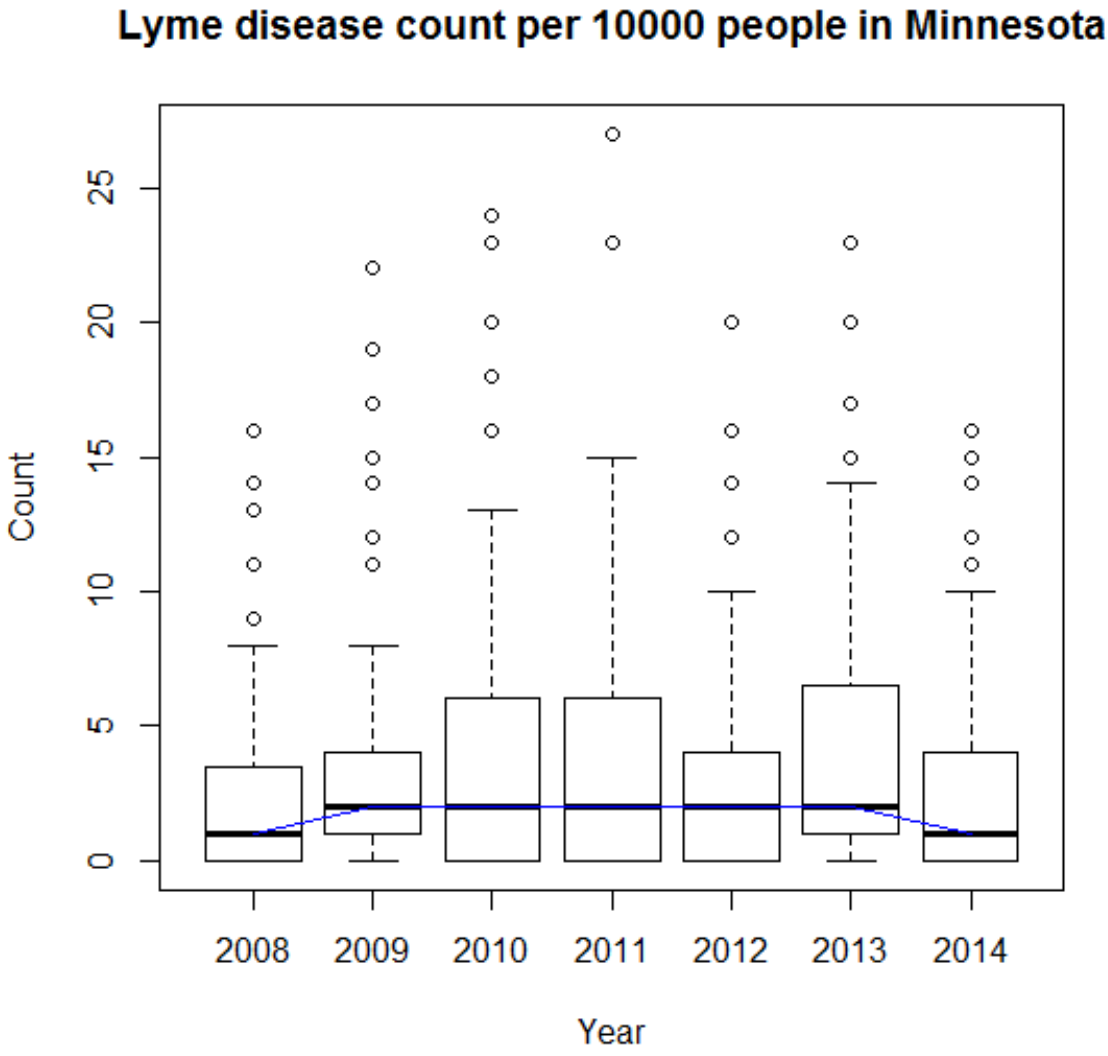


Figure 3. 5 Box and whisker plot of Lyme disease count in Minnesota from year 2012-2014.

The results from figure 3.6 and 3.7 show that the distribution of Lyme disease was clustered in counties in the northeastern part of the state of Minnesota.

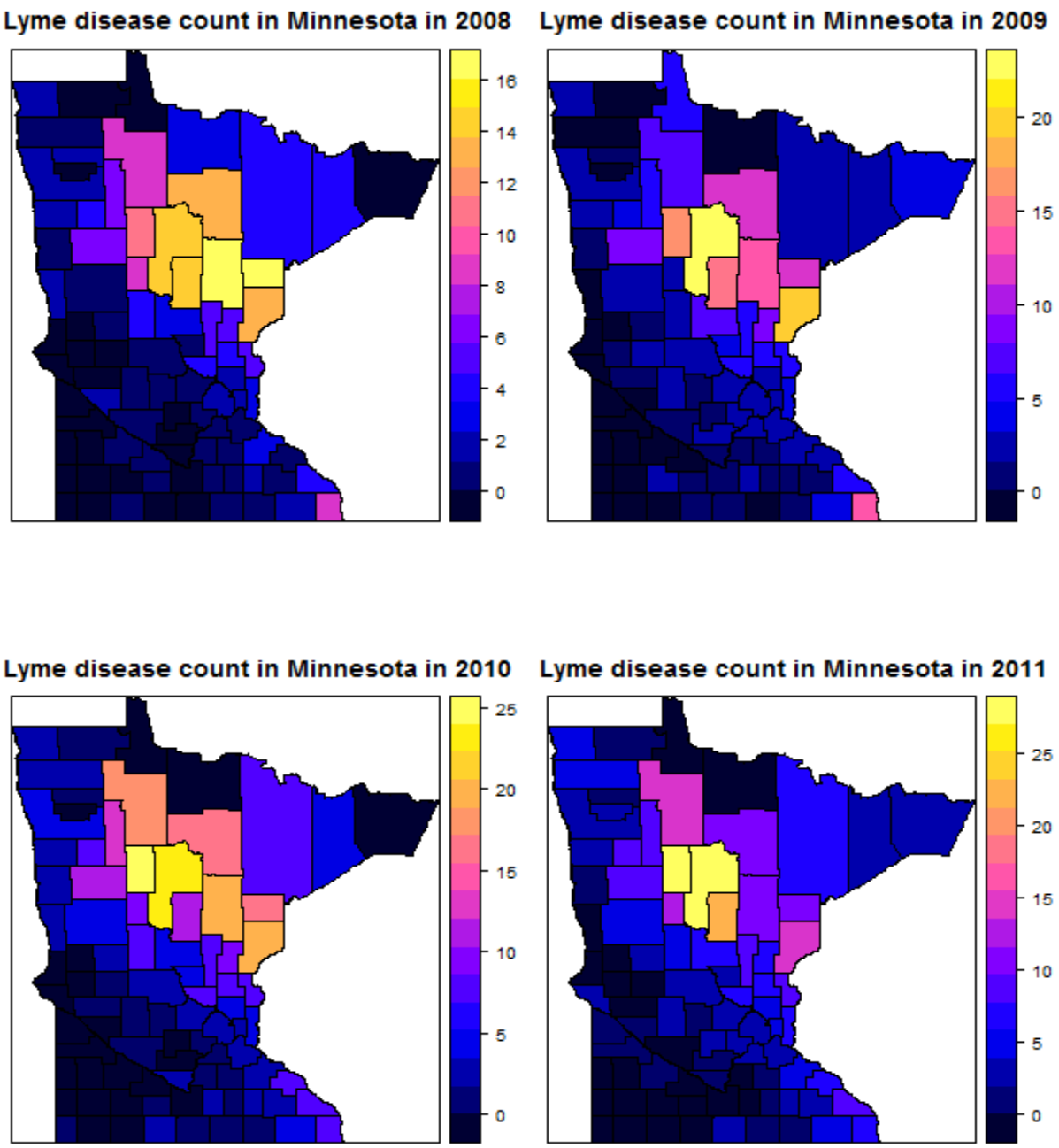


Figure 3. 6 County level map of Lyme disease count per 10000 people in Minnesota from year 2008-2011.

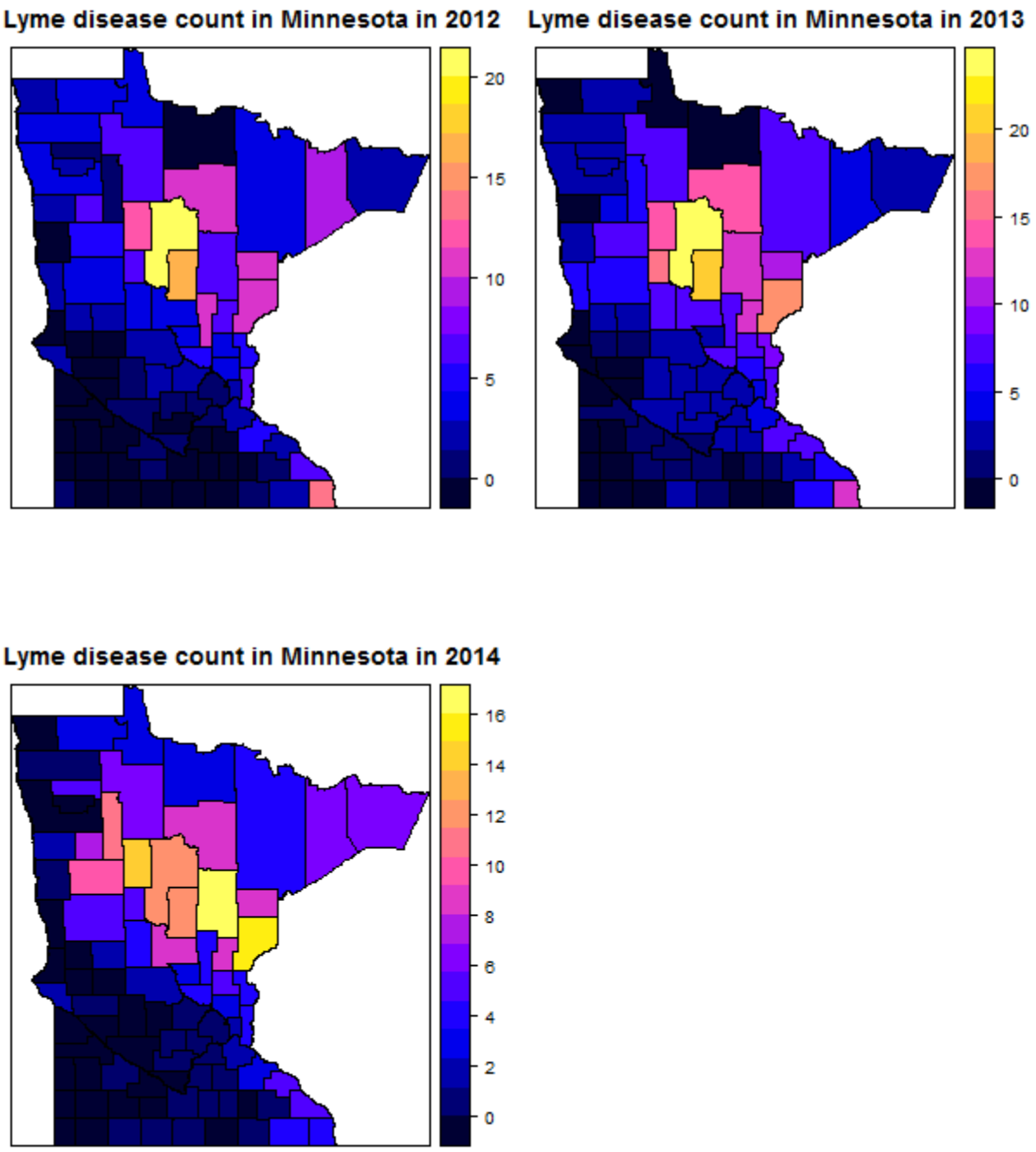


Figure 3. 7 County level map of Lyme disease count per 10000 people in Minnesota from year 2012-2014.

The results from figure 3.8 shows the pattern in neighborhood structure and which helps to identify more isolated or more central counties.

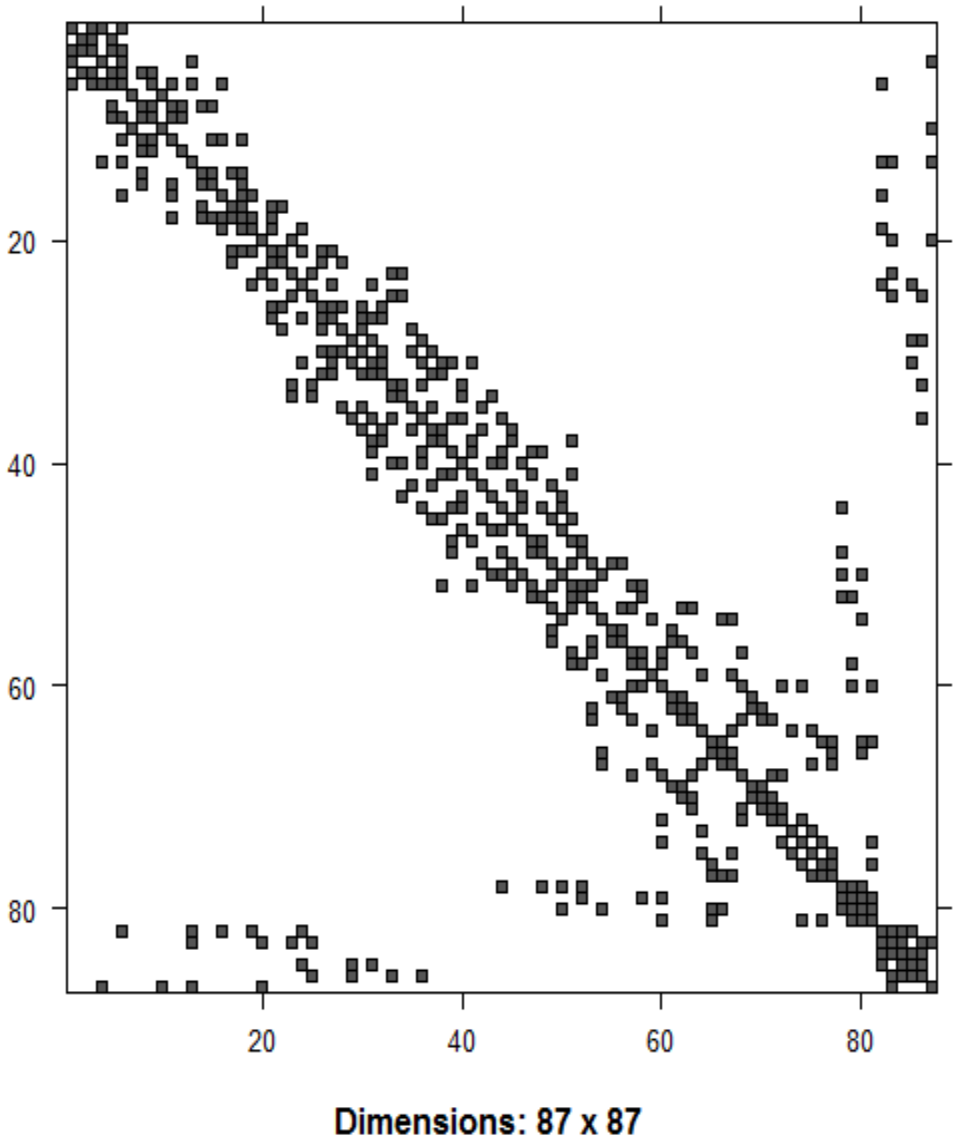


Figure 3. 8 Adjacency matrix of Minnesota counties.

Table 3.1 shows model coefficients from Poisson, Poisson hurdle (Zero inflated Poisson0) and Zero inflated Poisson (Zero inflated Poission1) models where β_0 is for intercept, β_1 for average annual temperature 2 years lag, β_2 for average annual precipitation two years lag,

β_3 for average winter temperature one year lag, β_4 for average winter temperature, β_5 for average annual temperature one year lag, β_6 for average annual temperature, β_7 for average annual precipitation, β_8 for average annual precipitation one year lag. The result also shows that there was significant relationship between Lyme disease cases and average annual temperature two years lag (β_1), average winter temperature (β_4), and average annual temperature (β_6). The effect of rest of the variables was not significant because 95% credible interval includes both positive and negative values. The result also shows that there was negative relationship between Lyme disease count and average annual temperature two years lag and average temperature whereas average winter temperature was positively related with Lyme disease count.

Table 3. 1: Model coefficients from Poisson, Poisson hurdle (Zero inflated Poisson0) and Zero inflated Poisson (Zero inflated Poission1) models.

Model	Covariates	Mean	SD	2.5 % quantile	0.975% quantile
Poisson	β0	10.009	2.349	5.398	14.594
	β1	-0.116	0.04	-0.193	-0.043
	β2	0.005	0.004	-0.002	0.011
	β3	-0.008	0.023	-0.051	0.037
	β4	0.061	0.018	0.028	0.097
	β5	-0.073	0.048	-0.17	0.019
	β6	-0.085	0.036	-0.154	-0.011
	β7	0.006	0.003	-0.001	0.013
	β8	0.003	0.003	-0.003	0.009
Zero inflated Poison0	β0	9.638	2.287	5.054	14.018
	β1	-0.114	0.041	-0.193	-0.035
	β2	0.005	0.004	-0.002	0.012
	β3	-0.007	0.023	-0.051	0.039
	β4	0.059	0.018	0.025	0.097
	β5	-0.077	0.051	-0.179	0.017
	β6	-0.068	0.038	-0.14	0.008
	β7	0.007	0.004	0.001	0.014
	β8	0.004	0.003	-0.003	0.01
Zero inflated Poisson1	β0	9.615	2.578	4.518	14.383
	β1	-0.109	0.043	-0.184	-0.028
	β2	0.005	0.003	-0.002	0.012
	β3	-0.009	0.023	-0.052	0.037
	β4	0.059	0.019	0.026	0.096
	β5	-0.069	0.05	-0.17	0.022
	β6	-0.086	0.035	-0.153	-0.015
	β7	0.006	0.004	-0.003	0.01
	β8	0.004	0.003	-0.003	0.01

Table 3.2 shows model coefficients from Negative binomial, Negative binomial hurdle (Zero inflated Negative binomial 0) and Zero inflated Negative binomial (Zero inflated Negative binomial 1) models where β_0 is for intercept, β_1 for average annual temperature 2 years lag, β_2 for average annual precipitation two years lag, β_3 for average winter temperature one year lag, β_4 for average winter temperature, β_5 for average annual

temperature one year lag, β_6 for average annual temperature, β_7 for average annual precipitation, β_8 for average annual precipitation one year lag.

Table 3. 2 Model coefficients from Negative binomial, Negative binomial hurdle (Zero inflated Negative binomial 0) and Zero inflated Negative binomial (Zero inflated Negative binomial 1) models.

Model	Covariates	Mean	SD	2.5 quantile	0.975 quantile
Negativebinomial	β0	6.268	1.861	2.62	9.93
	β1	-0.033	0.015	-0.064	-0.003
	β2	0.005	0.005	-0.005	0.014
	β3	-0.006	0.021	-0.048	0.036
	β4	0.039	0.012	0.016	0.062
	β5	-0.055	0.037	-0.128	0.017
	β6	-0.095	0.034	-0.161	-0.029
	β7	0.001	0.005	-0.008	0.01
	β8	0.011	0.005	0.001	0.02
Zeroinflated negativebinomial0	β0	6.43	1.804	2.883	9.973
	β1	-0.035	0.016	-0.066	-0.005
	β2	0.005	0.005	-0.005	0.015
	β3	-0.006	0.022	-0.05	0.038
	β4	0.04	0.011	0.017	0.061
	β5	-0.055	0.038	-0.13	0.018
	β6	-0.093	0.034	-0.16	-0.027
	β7	0.002	0.005	-0.008	0.011
	β8	0.012	0.005	0.0026	0.022
Zeroinflated negativebinomial1	β0	6.236	1.846	2.604	9.855
	β1	-0.033	0.015	-0.063	-0.003
	β2	0.005	0.005	-0.005	0.014
	β3	-0.006	0.021	-0.048	0.036
	β4	0.039	0.012	0.016	0.061
	β5	-0.055	0.037	-0.127	0.017
	β6	-0.095	0.034	-0.161	-0.029
	β7	0.001	0.005	-0.008	0.01
	β8	0.011	0.005	0.001	0.02

Table 3.3 presents results of model diagnostics. Zero-inflated Poisson model is the best model because it has the lowest DIC value.

Table 3. 3 Model diagnostics.

Model	DIC
Poisson	2887.75
Zero-inflated Poisson0	3210.87
Zero-inflated Poisson1	2884.51
Negative binomail	2891.28
Zero-inflated negative binomial0	3205.65
Zero-inflated negative binomial1	2890.3

Figure 3.9 displays posterior density plot for intercept, Temp.vector (average annual temperature 2 years lag), Prep.vector (average annual precipitation two years lag), t1 (average winter temperature one year lag), t2 (average winter temperature), ty1 (average annual temperature one year lag), ty (average annual temperature), py (average annual precipitation), and py1 (average annual precipitation one year lag).

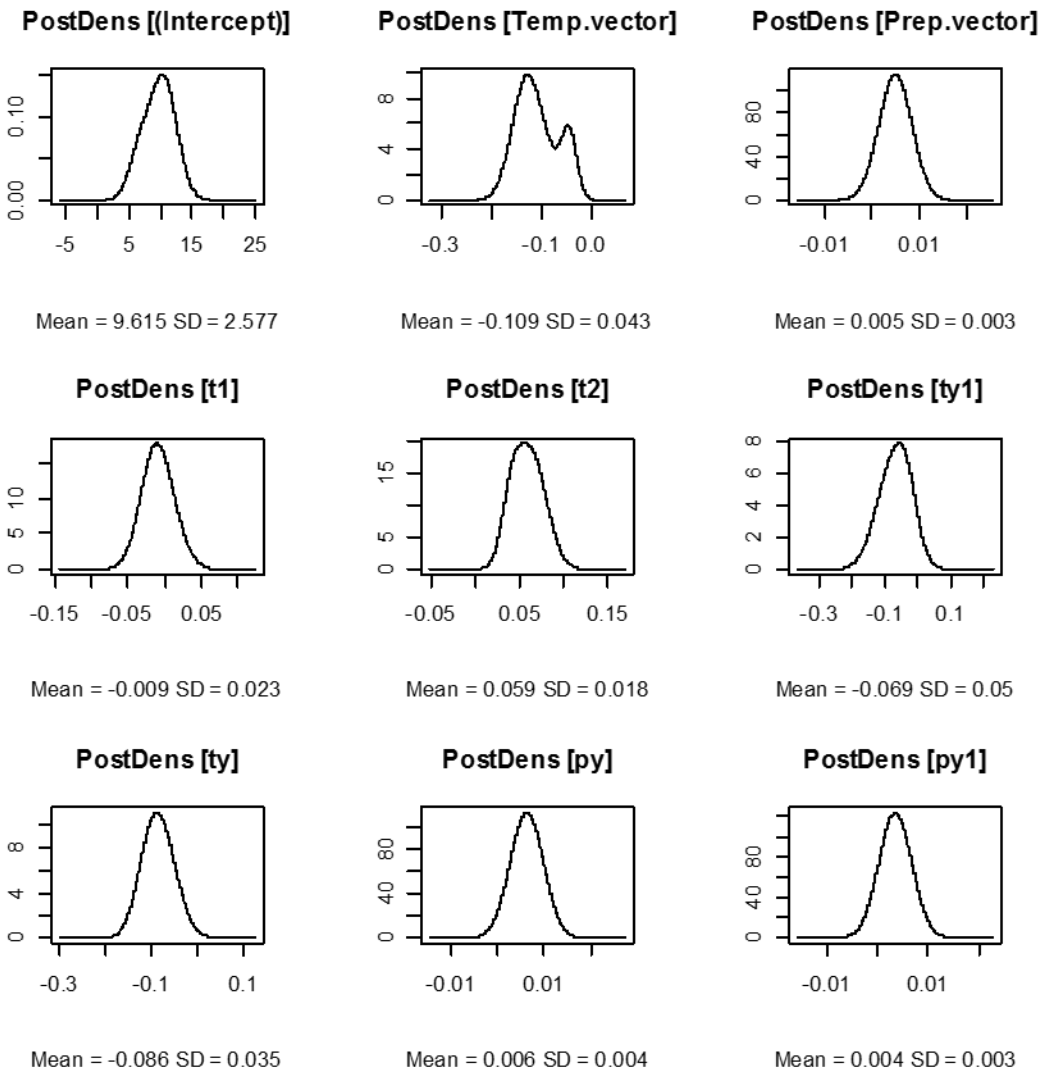


Figure 3. 9 Posterior density plots.

Figure 3.10 shows that there was no significant temporal trend of Lyme disease in Minnesota from 2008-2014.

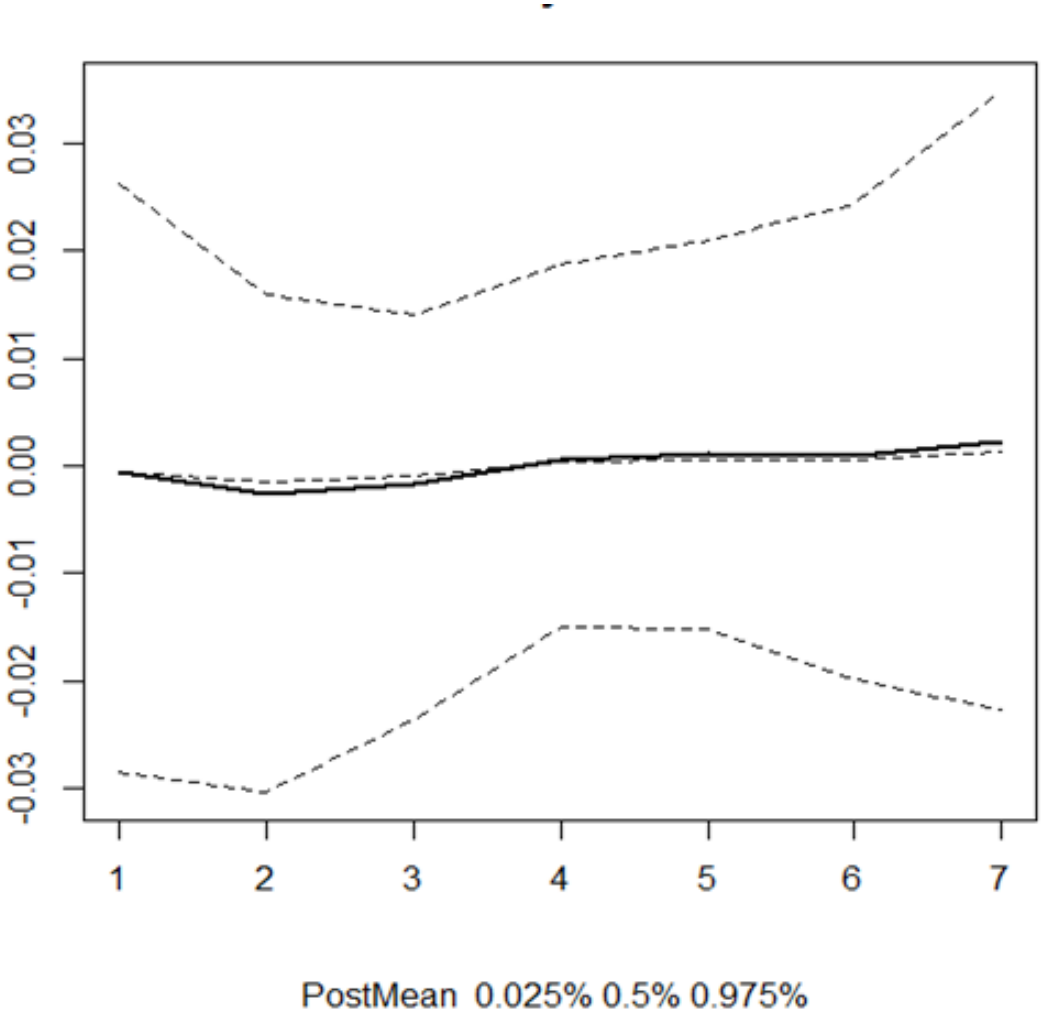


Figure 3. 10 Plot for posterior mean with 95 % credible interval over years.

Figure 3.11 displays diagnostic plots for zero-inflated Poisson regression. There is no failure in result means the model is good.

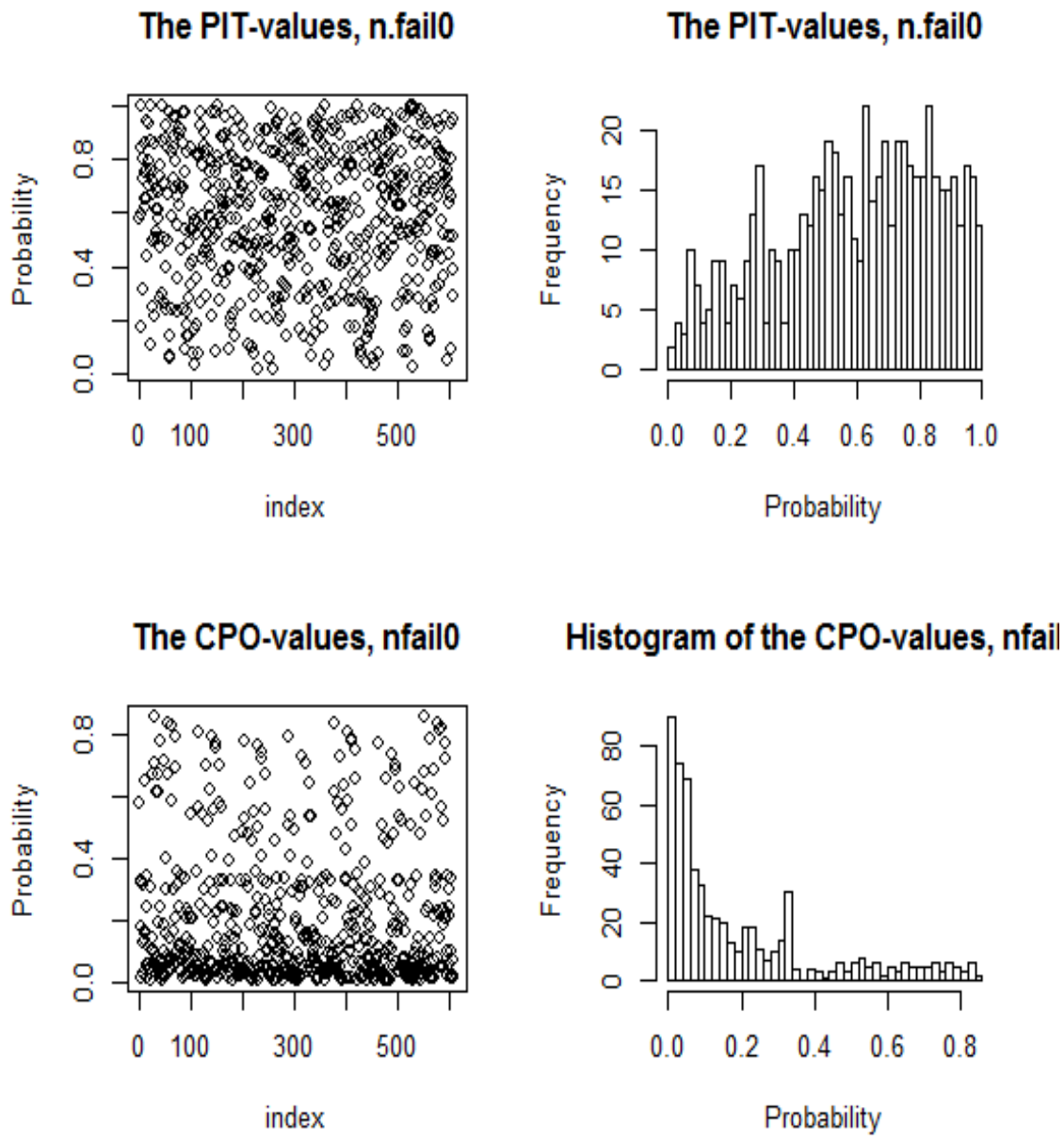


Figure 3. 11 Diagnostic plots.

3.4 Discussion and Conclusion

Spatiotemporal count data with excess zeros are very common in epidemiology. The modeling of spatiotemporal count data is more complicated due to the presence of excess zeros and spatiotemporal correlation. Researchers and statisticians have been using Bayesian hierarchical models to address those issues. Unfortunately, fitting Bayesian hierarchical model to estimate model parameters is complicated due to computational demands of MCMC algorithm. Recently developed INLA is increasingly popular as a substitute for MCMC methods to fit Bayesian hierarchical models due to less computational demand and accurate results. The use of INLA approach is increasingly popular in spatiotemporal count data analysis (DiMaggio, 2015; Musenge et al., 2013; Ross, Hooten, & Koons, 2012; Schrödle & Held, 2011; Serra, Saez, Juan, Varga, & Mateu, 2014; Ugarte, Adin, Goicoa, & Militino, 2014; Zhao et al., 2014).

The objective of this study was to study the relationship between Lyme disease count and climatic covariates in Minnesota. Like many other epidemiological data, the study of Lyme disease count regression model is complicated due to the presence of complicated statistical features such as excess of zeros and spatiotemporal correlation. Poisson and negative binomial models are the most commonly used models in count data modeling. The extension of these models is available to address the issue of presence of excess zeros. Regular and zero inflated Poisson and negative binomial models were fitted to study the relationship between Lyme disease count and climatic covariates using INLA approach. Zero-inflated Poisson model was selected as the best model based on DIC and effective numbers of parameters.

Among all the climatic variables tested in this analysis to study their association with Lyme disease count in Minnesota, average annual temperature two years lag, average winter temperature and average annual temperature were the only covariates that have the significant association with Lyme disease count. There was no clear temporal pattern of Lyme disease in Minnesota over the years 2008-2014. The findings of this study shows that average annual temperature two years lag and average annual temperature of the same year was negatively associated with Lyme disease. Similarly, average winter temperature of the same year was positively associated with Lyme disease which is consistent with previous studies by (Brownstein et al., 2005; Ostfeld, Canham, Oggenfuss, Winchcombe, & Keesing, 2006; Schauber, Ostfeld, & Evans Jr, 2005; Subak, 2003). The increase of average winter temperature helps to increase the activity of Lyme disease causing ticks in the summer months.

References

Agarwal, D. K., Gelfand, A. E., & Citron-Pousty, S. (2002). Zero-inflated models with application to spatial count data. *Environmental and Ecological Statistics, 9*(4), 341-355. doi:10.1023/a:1020910605990

Arab, A. (2015). Spatial and spatio-temporal models for modeling epidemiological data with excess zeros. *International journal of environmental research and public health, 12*(9), 10536-10548.

Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M., & Songini, M. (1995). Bayesian analysis of space—time variation in disease risk. *Statistics in Medicine, 14*(21-22), 2433-2443.

Best, N., Richardson, S., & Thomson, A. (2005). A comparison of Bayesian spatial models for disease mapping. *Statistical methods in medical research, 14*(1), 35-59.

Blangiardo, M., Cameletti, M., Baio, G., & Rue, H. (2013). Spatial and spatio-temporal models with R-INLA. *Spatial and spatio-temporal epidemiology, 7*, 39-55.

Brownstein, J. S., Holford, T. R., & Fish, D. (2005). Effect of Climate Change on Lyme Disease Risk in North America. *EcoHealth, 2*(1), 38-46. doi:10.1007/s10393-004-0139-x

Cressie, N., Calder, C. A., Clark, J. S., Hoef, J. M. V., & Wikle, C. K. (2009). Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. *Ecological Applications, 19*(3), 553-570.

DiMaggio, C. (2015). Small-area spatiotemporal analysis of pedestrian and bicyclist injuries in New York City. *Epidemiology, 26*(2), 247-254.

Eggers, J. (2015). On Statistical Methods for Zero-Inflated Models. UUDM Project Report 9.

Eisen, R. J., Eisen, L., & Beard, C. B. (2016). County-Scale Distribution of Ixodes scapularis and Ixodes pacificus (Acari: Ixodidae) in the Continental United States. *J Med Entomol, 53*(2), 349-386. doi:10.1093/jme/tjv237

Gage, K. L., Burkot, T. R., Eisen, R. J., & Hayes, E. B. (2008). Climate and vectorborne diseases. *American journal of preventive medicine, 35*(5), 436-450.

Johnson, T., Bjork, J., Neitzel, D., Dorr, F., Schiffman, E., & Eisen, R. (2016). Habitat suitability model for the distribution of Ixodes scapularis (Acari: Ixodidae) in Minnesota. *J Med Entomol, 53*(3), 598-606.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics, 34*(1), 1-14.

Lawson, A. B. (2013). *Bayesian disease mapping: hierarchical modeling in spatial epidemiology*: CRC press.

Li, J., Kolivras, K. N., Hong, Y., Duan, Y., Seukep, S. E., Prisley, S. P., . . . Gaines, D. N. (2014). Spatial and Temporal Emergence Pattern of Lyme Disease in Virginia. *The American Journal of Tropical Medicine and Hygiene, 91*(6), 1166-1172. doi:10.4269/ajtmh.13-0733

Lindsay, L. R., Barker, I. K., Surgeoner, G. A., McEwen, S. A., Gillespie, T. J., & Robinson, J. T. (1995). Survival and development of Ixodes scapularis (Acari: Ixodidae) under various climatic conditions in Ontario, Canada. *J Med Entomol, 32*(2), 143-152.

Martins, T. G., Simpson, D., Lindgren, F., & Rue, H. (2013). Bayesian computing with INLA: New features. *Computational Statistics & Data Analysis, 67,* 68-83. doi:http://dx.doi.org/10.1016/j.csda.2013.04.014

McCabe, G. J., & Bunnell, J. E. (2004). Precipitation and the occurrence of Lyme disease in the northeastern United States. *Vector-Borne and Zoonotic Diseases, 4*(2), 143-148.

Mead, P. S. (2015). Epidemiology of Lyme disease. *Infectious disease clinics of North America, 29*(2), 187-210.

Moore, S. M., Eisen, R. J., Monaghan, A., & Mead, P. (2014). Meteorological influences on the seasonality of Lyme disease in the United States. *The American Journal of Tropical Medicine and Hygiene, 90*(3), 486-496.

Musenge, E., Chirwa, T. F., Kahn, K., & Vounatsou, P. (2013). Bayesian analysis of zero inflated spatiotemporal HIV/TB child mortality data through the INLA and SPDE approaches: Applied to data observed between 1992 and 2010 in rural North East South Africa. *International Journal of Applied Earth Observation and Geoinformation, 22,* 86-98. doi:http://dx.doi.org/10.1016/j.jag.2012.04.001

Ogden, N., Lindsay, L., Beauchamp, G., Charron, D., Maarouf, A., O'Callaghan, C., . . . Barker, I. (2004). Investigation of relationships between temperature and developmental rates of tick Ixodes scapularis (Acari: Ixodidae) in the laboratory and field. *J Med Entomol, 41*(4), 622-633.

Ogden, N. H., St-Onge, L., Barker, I. K., Brazeau, S., Bigras-Poulin, M., Charron, D. F., . . . Maarouf, A. (2008). Risk maps for range expansion of the Lyme disease vector,

Ixodes scapularis, in Canada now and with climate change. *International Journal of Health Geographics, 7*(1), 24.

Ostfeld, R. S., Canham, C. D., Oggenfuss, K., Winchcombe, R. J., & Keesing, F. (2006). Climate, deer, rodents, and acorns as determinants of variation in Lyme-disease risk. *PLoS Biol, 4*(6), e145.

Ross, B. E., Hooten, M. B., & Koons, D. N. (2012). An accessible method for implementing hierarchical models with spatio-temporal abundance data. *PLoS One, 7*(11), e49395.

Rue, H., & Martino, S. (2007). Approximate Bayesian inference for hierarchical Gaussian Markov random field models. *Journal of statistical planning and inference, 137*(10), 3177-3192.

Schauber, E. M., Ostfeld, R. S., & Evans Jr, A. S. (2005). What is the best predictor of annual Lyme disease incidence: weather, mice, or acorns? *Ecological Applications, 15*(2), 575-586.

Schrödle, B., & Held, L. (2011). Spatio-temporal disease mapping using INLA. *Environmetrics, 22*(6), 725-734.

Serra, L., Saez, M., Juan, P., Varga, D., & Mateu, J. (2014). A spatio-temporal Poisson hurdle point process to model wildfires. *Stochastic environmental research and risk assessment, 28*(7), 1671-1684.

Stafford, K. C., 3rd. (1994). Survival of immature Ixodes scapularis (Acari: Ixodidae) at different relative humidities. *J Med Entomol, 31*(2), 310-314.

Subak, S. (2003). Effects of climate on variability in Lyme disease incidence in the northeastern United States. *American Journal of Epidemiology, 157*(6), 531-538.

Ugarte, M. D., Adin, A., Goicoa, T., & Militino, A. F. (2014). On fitting spatio-temporal disease mapping models using approximate Bayesian inference. *Statistical methods in medical research, 23*(6), 507-530.

Wikle, C. K., & Anderson, C. J. (2003). Climatological analysis of tornado report counts using a hierarchical Bayesian spatiotemporal model. *Journal of Geophysical Research: Atmospheres, 108*(D24).

Zhao, X., Cao, M., Feng, H.-H., Fan, H., Chen, F., Feng, Z., . . . Zhou, X.-H. (2014). Japanese encephalitis risk and contextual risk factors in Southwest China: A Bayesian hierarchical spatial and spatiotemporal analysis. *International journal of environmental research and public health, 11*(4), 4201-4217.

Chapter 4

4 Conclusions and future directions

This chapter includes major conclusions and contribution of this dissertation and suggests some possible areas for future research.

4.1 Conclusions

Spatial and spatiotemporal data analysis is a new and emerging field in statistics. Spatial and spatiotemporal regression models are quite common in epidemiological data analysis. Chapter one provides the general overview of spatial data and challenges of spatial data analysis. Chapter two summarize the models used for geographically weighted regression and fitted MGWR model to study the association between diabetes prevalence and socioeconomic and life style covariates. The benefit of fitting MGWR is that it can include both local and global variables in a single model. Chapter three presents the use of INLA approach in model selection for zero- inflated count data which is an alternative of computationally challenging MCMC method in Bayesian hierarchical modeling. INLA approach was used to find the best model for the spatiotemporal regression analysis to find the relationship of Lyme disease count data with climatic variables in Minnesota. Zero-inflated Poisson regression model was identified as the best model from the analysis.

4.2 Contributions

The research reported in chapter two serves as a new and improved analysis workflow for geographically weighted regression models. There are a large number of research articles

dealing with geographically weighted regression models, but there is no clear workflow for dealing such data. I propose the following workflow that will provide a clear guideline for applying geographically weighted regression models with spatial data.

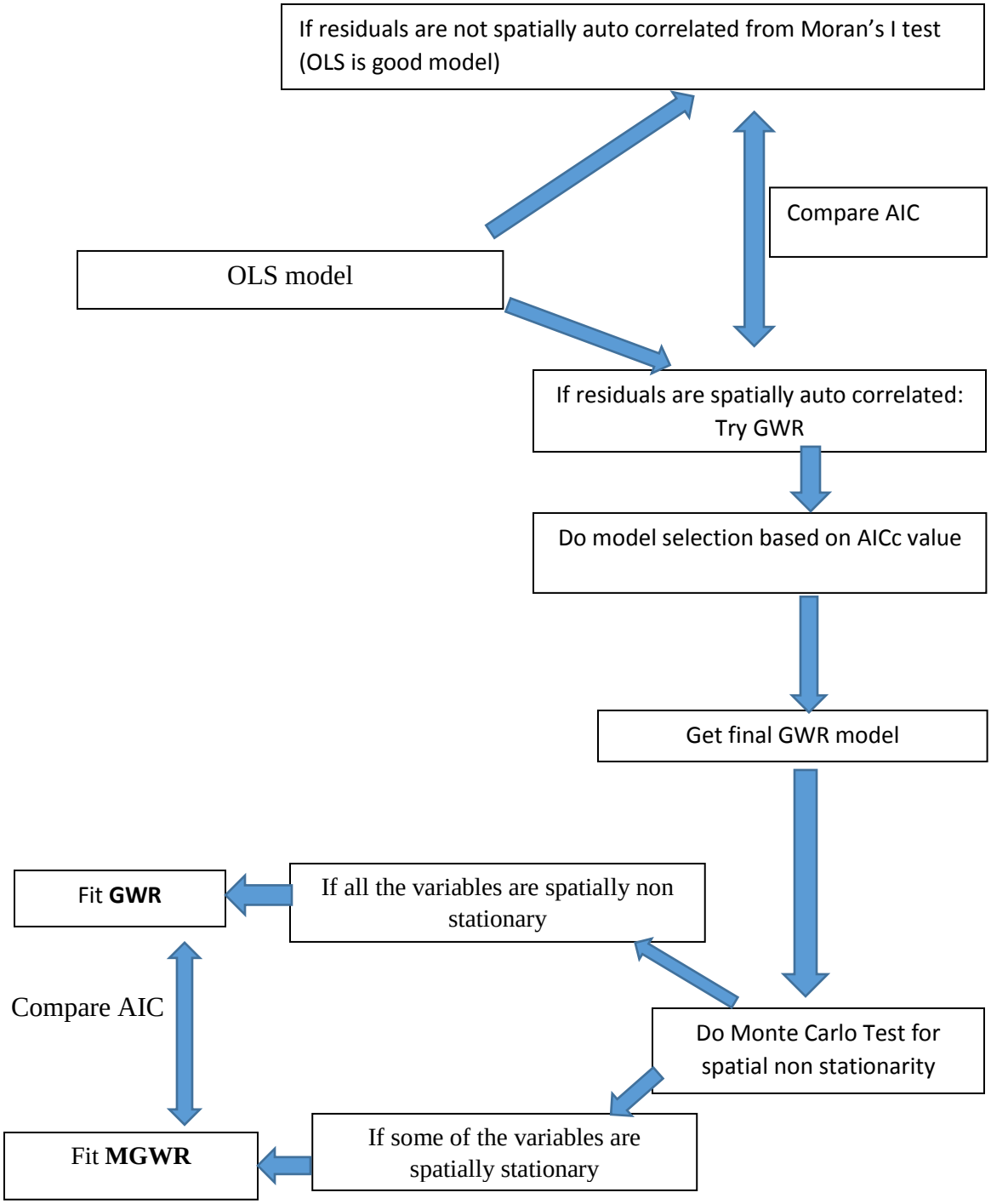


Figure 4. 2 Suggested flow chart for regression analysis of spatial data.

In addition, this study help to understand the association between diabetes prevalence and socioeconomic and life style factors by using MGWR which has never been considered for the analysis of any epidemiological data before. The findings of this study will help to understand the distribution of diabetes in the Midwestern United States and the effects of some covariates.

In chapter three, A Bayesian hierarchical model was developed using integrated INLA approach which runs faster than traditionally used MCMC methods for spatiotemporal data analysis. The use of INLA approach help to reduce the computational time so we can fit and compare the results of different models in time efficient way. The results from spatiotemporal regression analysis of the association between Study of the association between Lyme disease count data and climatic covariates in Minnesota by using zero-inflated spatiotemporal regression model helps to understand the effects of climatic variables on Lyme disease count. The findings of this study will help Lyme disease prevention and control.

4.3 Areas for future research

This research proposes an improved analysis workflow to fit geographically weighted regression models. Future work can be conducted in the study of other chronic diseases in different geographical scales. INLA approach is very fast in computation. Future work can be extended to regional to even national data of Lyme disease.