Spatio-Temporal Clustering of Lyme Disease in Texas: 2000-2017

Robert R. Wardrup

October 7, 2020

Abstract

To determine if and where Lyme disease (LD) spatio-temporal clusters occurred in Texas between 2000-2015 and if these clusters were correlated with deciduous forest patches, an analysis was done on the 254 counties of Texas. County LD counts were analyzed alongside deciduous forest patch attributes using Kendall's rank correlation to determine model inclusion suitability. SaTScan software was used for spatio-temporal cluster analysis. The results indicate that deciduous forest patches are not significantly correlated with LD rate. One statistically significant LD spatio-temporal cluster was detected in Southeastern Texas, highlighting an area with increasing rate and a need for greater resource allocation.

1 Introduction

Lyme borreliosis, hereafter referred to as Lyme disease (LD), is a vector-borne autoimmune disease that can have serious and long-lasting health consequences. It was first discovered in Connecticut in 1975 when a cluster of children and adults presented with unusual arthritic symptoms [27] and has since become the most commonly reported vector-borne disease in the United States with more than 20,000 cases reported each year [3, 24]. The disease quickly spread to other states in the Northern and Northeastern United States and has since become endemic to the area while continuing its westward expansion [4, 27]. The habitat suitability of the primary vector of LD, the arthropod *Ixodes Scapularis*, is expected to expand in Texas through 2050 before beginning a gradual recession through 2080 [6].

Persons infected with LD typically develop a slow-spreading rash 3-32 days after the infecting tick bite. Flu-like symptoms including fatigue, malaise, fever, and myalgias also typically present at this time. Following antibiotic treatment, a small percentage of patients develop persistent arthritis, facial palsy, heart palpitations and/or neurological difficulties [10, 27]. LD mainly affects children aged 5-9 years old and adults aged 45-55 years old [25].

Forest fragmentation is the result of natural and/or anthropomorphic processes which result in four outcomes: a continuous forest area becomes discontinuous, with two or more forest patches; loss of habitat; spatial patch isolation due to lack of habitat connectivity; change in habitat-matrix interactions [5]. Forest fragmentation may affect the spatial distribution of disease vectors while at the same time affecting the interaction between humans and wildlife [12]. The presence of infected vectors has been shown to have a positive relationship with forest fragmentation [1] while having either a positive [7] or negative [18] relationship with human LD rate.

The goal of this study is to determine if and when spatio-temporal clustering of Lyme disease rates have occurred in Texas and if forest patches are significant factors of the clustering effect. Such an understanding will highlight areas of Texas which may require further

intervention to halt trends of increasing LD rate that are not simply explained by the area's vegetation.

2 Materials and Methods

2.1 Study Area

Texas, the largest state in the contiguous United States, has a burgeoning estimated 2015 population of 27,695,284 [29], the majority of which lives in the Eastern half of the state. Texas espouses several climate regions and ecological niches which in turn gives opportunity for interaction between human, wildlife, and zoonotic diseases. There are 254 counties in the state, with a median 2010 population of 18,380 per county. Counties were chosen as the study geographies as county-level data is the most granular available from the CDC.

2.2 Case and Population Data

County-level LD case counts were obtained for the years 2000-2014 [8]. The CDC defines a case as a person displaying at least one of three markers: erythema migrans (EM), which occurs on 60% to 80% of cases and known exposure to *I. Scapularis*, EM with no known exposure but a positive lab test, or at least one late manifestation symptom and positive lab test [9, 28]. The CDC updated the LD case definition in 2009 following improvements in testing and case confirmation techniques. The case count saw a spike during that year before the number returned to previous levels the following year [3].

Population data at the county level was obtained from the United States Census Bureau and merged with case count data in a geographical information system (GIS) shapefile. The mean population for the 2000 and 2010 census was calculated for each county to provide a reasonable estimate of the 2005 population [28].

2.3 NDVI

Both the spatial distribution of LD rate and the presence of *I. Scapularis* has been shown to be positively correlated with a location's NDVI in the spring and fall, when the greatest contrast between forested and non-forested areas is noted [20]. Taking an area's vegetation into account would allow for the discovery of clusters that are not present due to only ecological factors. The vegetation of study areas is measurable through satellite imagery by the Normalized Difference Vegetation Index (NDVI), which is obtained via the formula [14]:

$$\text{NDVI} = \frac{NIR - VIS}{NIR + VIS}, \tag{1}$$

where NIR and VIS are the near-infrared and visible wavelengths, respectively. NDVI values range from -1, representing non-vegetation such as lakes, rivers and oceans, to 1, representing dense forestry. In order to detect spatio-temporal clusters of LD that are not affected by the area's environment greenness, NDVI was used as a co-variant in the analysis.

NDVI composites were obtained from the United States Geological Survey's Advanced Very High Resolution Radiometer (AVHRR) for mid-April for years 2000-2014. AVHRR data is provided by the USGS in 8-bit binary format which does not allow for decimal values. As such, the USGS provides rescaled NDVI data to a range of 0-200. These values were rescaled to the standard NDVI range by using the equation provided by the USGS (personal communication, 8 December, 2015):

$$\text{NDVI} = (x/100) - 1, \tag{2}$$

before non-vegetative values were removed. Because the resultant NDVI values were skewed, the median pixel value per county was used in lieu of the mean. The median pixel values were then categorized for SaTScan using quantiles, such that an equal number of pixels fell into each of 5 categories.

2.4 Forest Patches

Forest patches were derived from data provided by the National Land Cover Database (NLCD) [16]. The raster data is provided at a 30-meter resolution for the contiguous United States. This data was masked by a shapefile representing the state boundary of Texas before it was reprojected into the USA Contiguous Equidistant Conic geographic projection. The forest pixels were measured within a 3x3 pixel moving window, which results in a moving window that is 90m x 90m with an area of 8,100 square meters. From the data acquired from the NLCD, pixels which represented deciduous forest areas were extracted for further analysis. Prior studies have shown that deciduous forests are significantly correlated with *I. Scapularis* abundance [13, 19].

This method categorizes forest pixels based on two coefficients: pf and pff. Pf is the proportion of the number of forest pixels to the number of non forest pixels within the window. Pff equals the proportion of adjacent pixel pairs for which both pixels represent forest to the number of pixel pairs for which one pixel represents forest. Using this method, six categories were generated: interior forests (pf $= 1.0$), patch (pf < 0.4), transitional ($0.4 <$ pf < 0.6), edge (pf > 0.6 and pf - pff > 0), perforated (pf > 0.6 and pf - pff < 0) and undetermined, where pf > 0.6 and pf $=$ pff. Due to the moving window parameters, forest patches are defined as forest pixels which share a 8,100 square meter area with two or less pixels [26].

For this study, only forest pixels representing forest patches were used. Because forest patches frequently lie across county boundaries, the pixels were converted to vector data before they were clipped by the Texas county administrative boundaries. The area and distance to the nearest neighbor were calculated for each of the resultant polygons. The median patch area and distance to nearest neighbor were calculated for each County before the observations were categorized using the quantile method.

2.5 Statistical Analysis

Study regions with small populations may exhibit unstable variances in disease rates due to the effect of the small denominator. In order to address the issue, the empirical Bayes spatial smoothing method was used, which is a Bayesian approach to address the small-numbers problem by taking the observations of the entire state into account, reducing the inflation of rate in counties with low population. The smoothed rates were used in the Kendall correlation and the Moran's I tests.

Counties with a large population at risk experience little to no effect on their rate. Counties with a small population at risk, and thus a large variance, have their rates shrunk toward the global mean [2]. This method has been used in various other studies on disease rate. [28] similarly employed the method in studying LD temporal clusters in Texas. [17] used empirical Bayes smoothing in studying dengue fever infections in Australia. Finally, [23] used the method to study Chagas disease in Brazil.

Kendall's tau rank correlation was used to detect relationships between the smoothed LD rate, median NDVI, median forest patch area, and median distance between forest patches. Kendall's tau was chosen as there were ties in the data (occurrences of identical data points) due to the fact that multiple county's LD rate remained the same value across multiple years, and the test requires no assumptions of the distribution of the data. The rank correlation test outcomes were used to determine the inclusion or exclusion of the variables as co-variates in the spatial scan statistic.

To test whether or not LD rates were spatially auto-correlated during any of the years of the study period, Moran's I was calculated for each year using 9,999 Monte-Carlo replications. Moran's I is an indicator of a variable's correlation with itself in space. Moran's I ranges from -1, where the variable is perfectly dispersed, to 1, where the variable is perfectly auto-correlated. The test operates under the null hypothesis that there is no spatial auto-correlation in the variable. To reduce the chance of type-1 errors due to multiple testing, the resultant

p-values were then corrected using the Holm-Bonferroni method [15]. The Holm-Bonferroni method is a P-Value correction that takes into account the number of tests. The method addresses issues where the simple Bonferroni adjustment is too conservative.

Moran's I was calculated using the inverse-distance weighting (IDW) method of neighbor determination. IDW gives points which are closer together a higher weight than points which are further apart. In this instance, IDW holds that counties with high LD rates should be surrounded by counties with similar rates and vice-versa.

For the analysis, the R package rsatscan was used in lieu of SaTScan's graphical user interface. The rsatscan package allows advanced statistical analysis of SaTScan's output, and facilitates visualization of the data. A retrospective spatial variations in temporal trends analysis, using a Poisson probability model was conducted to determine if, when, and where statistically significant variations in LD rate trends were occurring.

SaTScan works by imposing an expanding scanning window onto the map and calculating the temporal tremd both inside and outside of the window to determine if the rate inside of the window is significantly higher than expected. The window is moved across the entire study area and temporal trends are calculated both inside and outside the window. Significance is determined by calculating a likelihood ratio which is higher the more unlikely it is that the difference in trends is due to pure chance. The output statistics include Relative Risk (RR), which is defined as the ratio of the probability that a resident of a county located within a cluster develops LD versus the probability that a resident of a non-cluster county develops LD. A person living in a county with a RR of 1, for example, is equally as likely to develop LD as a person living in a non-cluster county.

The test searches for a primary cluster (that which is least likely to have occurred by chance), and multiple secondary clusters. SaTScan allows the configuration of the maximum population at risk, so that the researcher may bias the analysis in search of smaller clusters. For this study, the maximum population at risk was set to 50% of the Texas population in order to prevent overlooking smaller clusters with no geographical overlap was allowed between the

clusters. Finally, 999 Monte-Carlo simulations were run in order to obtain greater statistical power [11, 21, 22].

3 Results

The analysis revealed that there were a total of 1,488 LD cases in Texas between 2000 and 2014. There were 0.4 annual cases per 100,000 with an overall 0.043% annual increase in the overall rate. The number of cases per year, per county ranged from zero cases in the majority of the counties to 47 cases in 2009 in Dallas county (Table 1).

Table 1: 2000-2014 Texas Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Median	Max
Population	3,810	92,236	329,837	67	17,867	4,092,459
Cases	3,810	0.39	1.97	0	0	47
Smooth Rate	3,810	0.46	1.01	0.00	0.31	44.89
NDVI	3,810	0.48	0.16	0.09	0.49	0.82
Patch Area	3,810	885.83	155.05	0	900	1,800
Patch Distance	3,810	312.71	4,809.07	0.00	42.43	85,856.50

3.1 Kendall's Tau and Moran's I

When smoothed LD rate was tested for correlation with median NDVI, median patch area, and median distance to patch nearest neighbor (NN), none of the results were significant at $\alpha = 0.05$ (See Table 2). Thus, the null hypothesis was not rejected: there is insufficient evidence to support the hypothesis that median NDVI, median patch area, and median distance to NN are correlated to LD rate.

The Monte-Carlo permutation of Moran's I showed that for only e out of the 15 years of study, there was statistically significant auto-correlation, as shown in Table 3.

Table 2: Results of Kendall's Rank Correlation

Variable	Z	P	τ
Median NDVI	-3.00	0.77	0.00
Median Patch Area	0.67	0.50	0.01
Median Patch NN Distance	-1.48	0.13	0.02

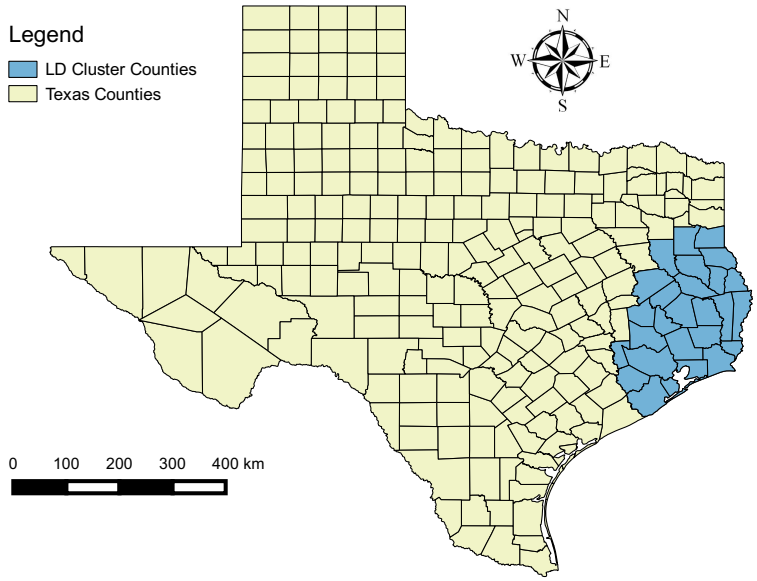
Table 3: Results of Monte-Carlo Simulations of Moran's I for the study period 2000-2014 for all Texas

Year	I	P	$adj.P$
2000	0.0657	0.040	0.360
2001	0.0619	0.045	0.362
2002	0.0639	0.046	0.362
2003	0.0969	0.013	0.157
2004	0.1831	0.001	0.010**
2005	0.0852	0.021	0.212
2006	0.0255	0.112	0.561
2007	0.0382	0.122	0.561
2008	0.0611	0.051	0.362
2009	0.3685	0.001	0.002**
2010	0.0854	0.015	0.162
2011	-0.0179	0.421	0.637
2012	0.1380	0.002	0.030*
2013	0.0060	0.262	0.637
2014	0.0192	0.212	0.637

Note: *and **denote significance at 0.05 and 0.01 respectively.

3.2 SaTScan Cluster Analysis

Figure 1: LD Counties included in cluster($P = 0.001$).



The SaTScan analysis revealed one statistically significant ($P < 0.05$) cluster (Figure 1) that encompassed a total of 27 counties, 16 of which exhibited growing LD rates, 1 of which exhibited a negative trend, and the remaining of which exhibited no detectable trend (where case counts were typically zero) in the Southeastern portion of Texas ($LLR = 29.27$, $P = 0.001$).This cluster exhibited a strong positive trend at 9.252%, while the counties outside that cluster (with that cluster removed) trended negatively at 2.43%. This cluster had a RR of 0.67, indicating that those living within this cluster are 0.67 times as likely to contract LD versus those who do not live in the cluster.

Table 4: Summary Statistics for Southeastern Cluster Detected by SaTScan

Statistic	N	Mean	St. Dev.	Min	Median	Max
Local Cases	27	11.00	29.10	0	2	145
Local Expected	27	14.93	46.15	0.57	3.13	242.36
Local Observed / Expected	27	0.85	0.78	0.00	0.67	2.61
Local Relative Risk	27	0.85	0.78	0.00	0.67	2.63
Local Trend	17	13.77	8.53	−1.91	12.66	25.64

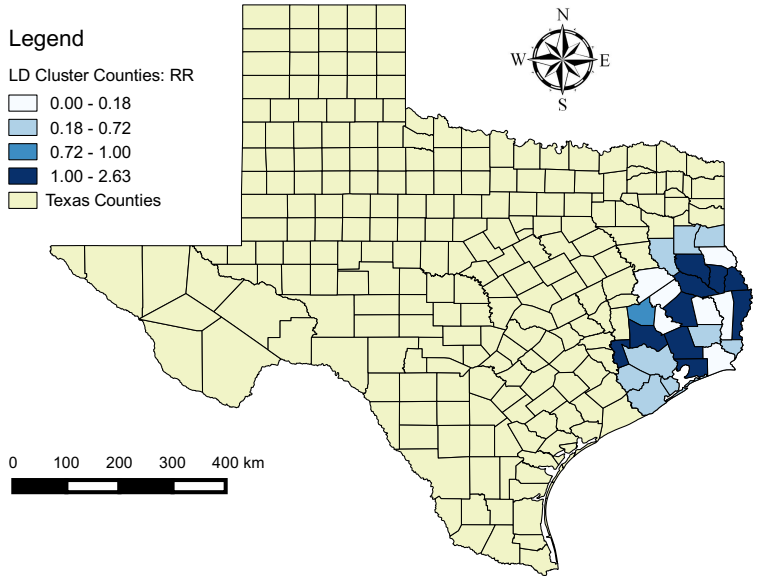


Figure 2: Local Lyme Disease RR

4 Discussion

While a previous study was successful in detecting purely spatial clusters of LD in Texas [28], this study is the first to detect a cluster of LD in Texas in the space-time realm. A statistically significant cluster of positive-trended LD rate was located in the Southeastern portion of Texas.

The results of the Kendall's rank correlation test show that there are no statistically significant correlations between LD rate, mean forest patch area and mean distance between

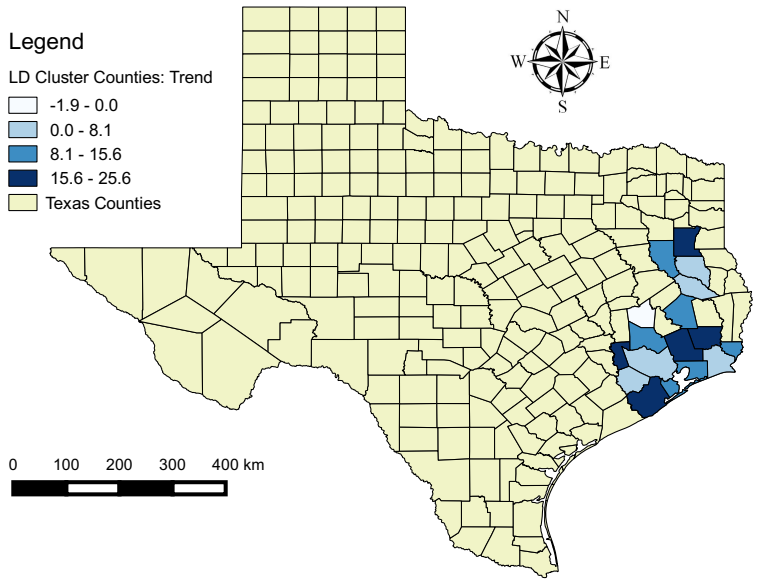


Figure 3: Local Lyme Disease Trends

forest patch NN. These results corroborate previous research which indicated that, while environmental variables are correlated with *b. Burgdorferi* infection of *I. Scapularis*, the variables are not necessarily correlated with LD rate [3].

The clustering of increasing LD rates raises questions of the efficacy of intervention measures in those clusters, and what can be changed in order to more effectively intervene so that the trend can be broken. While the Southeastern counties were not found to be significant spatial clusters in previous studies, the spatio-temporal method was able to detect clustering of counties in the region whose rates were increasing during the study period. This possibly indicates that the spatio-temporal cluster may develop into a spatial cluster in the near future.

As was noted in the Szonyi study, LD counts could be problematic as the rate of false-positives may be high due to the confusion of some physicians between LD and the Southern Tick-Associated Rash Illness (STARI). The two diseases present similar symptoms and thus

cases reported as LD could simply be incorrectly reported STARI cases, obfuscating studies which are based on case count data. This itself is an issue that must be addressed by studies, policymakers and physicians in the future.

Another potential confounding factor is the presence, or lack of presence, of so-called "Lyme-literate doctors," doctors who have training in LD treatment. The presence of these physicians could create artificial LD hot spots as patients travel to the doctor's locale, seeking treatment.

These results could help guide public help officials identify policies that would be effective intervention in LD trends in these counties.

5 Conclusion

This study has shown that there are significant spatio-temporal trends of LD rates occurring in Southeastern Texas. While previous studies have not shown spatial clusters in Southeastern Texas, the presence of a high-rate spatio-temporal cluster indicates that the area may become a spatial cluster in the future.

The presence of spatio-temporal clusters highlight areas that may need intervention in the near future in order to end the trend of increasing rates and to prevent LD from becoming endemic to Texas. Future studies are needed in order to determine if the presence of spatial and spatio-temporal clusters are related to the location of physicians who are "Lyme literate."

References

[1] B. F. Allan, F. Keesing, and R. S. Ostfeld. Effect of forest fragmentation on lyme disease risk. *Conservation Biology*, 17(1):267–272, 2003.

[2] L. Anselin, Y. W. Kim, and I. Syabri. Web-based analytical-tools for the exploration of spatial data. *Journal of Geographical Systems*, 6:197–218, 2004.

[3] S. F. Atkinson, S. Sarkar, A. Aviña, J. A. Schuermann, and P. Williamson. A determination of the spatial concordance between Lyme disease incidence and habitat probability of its primary vector Ixodes scapularis (black-legged tick). *Geospatial Health*, 9(1):203–212, 2014.

[4] a. G. Barbour and D. Fish. The biological and social phenomenon of Lyme disease. *Science (New York, N.Y.)*, 260(5114):1610–6, jun 1993.

[5] J. Bogaert, Y. S. S. Barima, L. I. W. Mongo, I. Bamba, A. Mama, M. Toyi, and R. Lafortezza. Forest Fragmentation: Causes, Ecological Impacts and Implications for Landscape Management. *Landscape Ecology in Forest Management and Conservation*, pages 273–296, 2011.

[6] J. S. Brownstein, T. R. Holford, and D. Fish. Effect of Climate Change on Lyme Disease Risk in North America. *EcoHealth*, 2(1):38–46, mar 2005.

[7] J. S. Brownstein, D. K. Skelly, T. R. Holford, and D. Fish. Forest fragmentation predicts local scale heterogeneity of Lyme disease risk. *Oecologia*, 146:469–475, 2005.

[8] Center for Disease Control & Prevention. Data and Statistics | Lyme Disease, 2015.

[9] Centers for Disease Control & Prevention. Lyme Disease 2011 Case definition, 2011.

[10] Centers for Disease Control & Prevention. Signs and Symptoms | Lyme Disease | CDC, 2015.

[11] M. Dwass. Modified Randomization Tests for Nonparametric Hypotheses. *The Annals of Mathematical Statistics*, 28(1):181–187, 1957.

[12] A. Estrada-Peña, R. S. Ostfeld, A. T. Peterson, R. Poulin, and J. de la Fuente. Effects of environmental change on zoonotic disease risk: an ecological primer. *Trends in Parasitology*, 30(4):205–214, 2014.

[13] M. Guerra, E. Walker, C. Jones, S. Paskewitz, M. R. Cortinas, A. Stancil, L. Beck, M. Bobo, and U. Kitron. Predicting the risk of Lyme disease: habitat suitability for Ixodes scapularis in the north central United States. *Emerging infectious diseases*, 8(3):289–97, 2002.

[14] J. W. Herring and David. Measuring Vegetation (NDVI & EVI), 2000.

[15] S. Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[16] C. Homer, J. Dewitz, L. Yang, S. Jin, P. Danielson, G. Xian, J. Coulston, N. Herold, J. Wickham, and K. Megown. Completion of the 2011 National Land Cover Database for the conterminous United States-Representing a decade of land cover change information, 2015.

[17] W. HU, A. CLEMENTS, G. WILLIAMS, and S. TONG. Spatial analysis of notified dengue fever infections. *Epidemiology and Infection*, 139(03):391–399, mar 2011.

[18] L. E. Jackson. Towards landscape design guidelines for reducing Lyme disease risk. *International Journal of Epidemiology*, 35(2):315–322, 2005.

[19] U. Kitron, J. K. Bouseman, and C. J. Jones. Use of the ARC/INFO GIS to study the distribution of Lyme disease ticks in an Illinois county. *Preventive Veterinary Medicine*, 11(3-4):243–248, dec 1991.

[20] U. Kitron and J. J. Kazmierczak. Spatial Analysis of the Distribution of Lyme Disease in Wisconsin. *American Journal of Epidemiology*, 145(6):558–566, 1997.

[21] M. Kulldorff. A spatial scan statistic, 1997.

[22] M. Kulldorff, L. Huang, and K. Konty. A scan statistic for continuous data based on the normal probability model. *International journal of health geographics*, 8:58, 2009.

[23] F. R. Martins-Melo, A. Novaes Ramos, C. H. Alencar, W. Lange, and J. Heukelbach. Mortality of Chagas' disease in Brazil: Spatial patterns and definition of high-risk areas. *Tropical Medicine and International Health*, 17(9):1066–1075, 2012.

[24] K. P. Messier, L. E. Jackson, J. L. White, and E. D. Hilborn. Landscape risk factors for Lyme disease in the eastern broadleaf forest province of the Hudson River valley and the effect of explanatory data classification resolution. *Spatial and Spatio-temporal Epidemiology*, 12:9–17, 2015.

[25] G. a. Poland. Prevention of Lyme disease: a review of the evidence. *Mayo Clinic proceedings. Mayo Clinic*, 76(7):713–724, 2001.

[26] K. Riitters, J. Wickham, R. O'Neill, B. Jones, and E. Smith. Conservation Ecology: Global Scale Patterns of Forest Fragmentation, 2000.

[27] A. C. Steere, J. Coburn, and L. Glickstein. Review series The emergence of Lyme disease. 113(8):1093–1101, 2004.

[28] B. Szonyi, I. Srinath, M. Esteve-Gassent, B. Lupiani, and R. Ivanek. Exploratory spatial analysis of Lyme disease in Texas –what can we learn from the reported cases? *BMC Public Health*, 2015.

[29] TXDSHS. Population Data (Projections) for Texas Counties, 2015, 2015.