Using Remote Sensing and Climate Data to Predict Relative Numbers of Lyme Disease Cases in the Northeastern United States

Jack Hester[a]

[a] *The Brown University School of Public Health*

Abstract

Lyme disease affects hundreds of thousands of Americans each year. While treatable if caught in time, the disease is associated with massive annual healthcare costs. If public health professionals and other researchers are able to determine which geographical regions are likely to see the largest number of Lyme disease cases during peak Lyme disease season, they can better target their preventative and monitoring resources, and ideally reduce the number of infections and associated medical costs. When coupled with a priori knowledge of tick behavior and Lyme disease transmission, remote sensing and climate data can provide information to assist in these geographical predictions. Lyme disease cases by county in 13 northeastern U.S. states for each year from 2000-2018 (n=4208) were compared with average temperature, average precipitation, and average leaf area index for the same counties in the months leading up to the heaviest part of the Lyme disease season. The results of this analysis suggest that these remote sensing and climate data sources are helpful in predicting regions that will see high numbers of Lyme disease cases in future months, and that remote sensing data is likely worth considering to predict health outcomes more broadly.

Keywords: Lyme disease, remote sensing, climate data

1. Introduction

It is estimated that 300,000 people are infected with Lyme disease each year in the United States alone.[1] If not treated quickly with antibiotics, Lyme disease can cause heart, nervous system, and joint problems, and even those who are treated often experience moderate flu-like symptoms. Beyond the burden on individual health, the economic burden of Lyme disease in the U.S. is estimated to be between $712 million and $1.3 billion per year.[2]

Within the United States, Lyme disease is most common in the northeastern states.[1] Cases occur year-around, but the most cases are seen from May through August.[3]

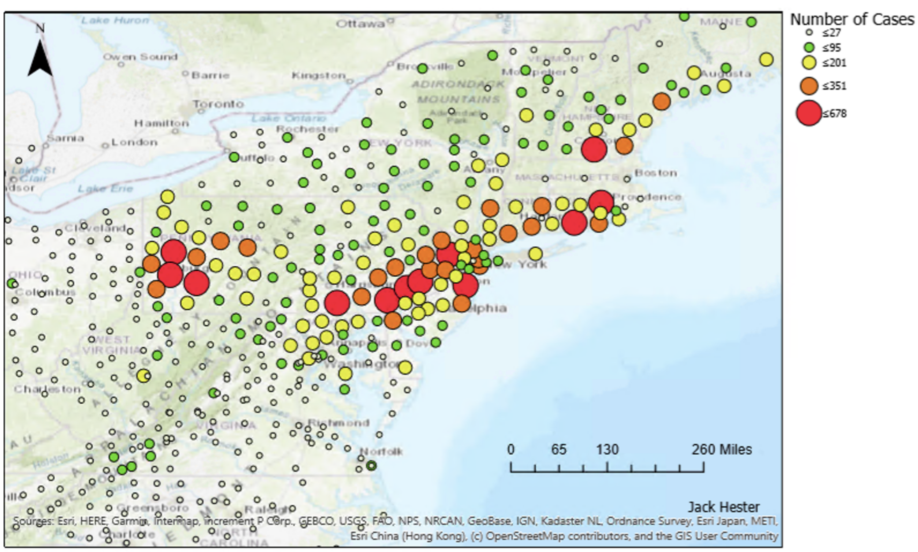

Figure 1: Distribution of reported Lyme disease by county in 2018 in the northeastern United States

Lyme disease is a bacterial infection spread by the deer tick (Ixodes scapularis).[4] These ticks prefer wooded or grassy areas, as well as humid regions (humidity greater than 85%) with temperatures over 45 degrees Fahrenheit.[5] As the name suggests, they also frequently parasitize deer, especially white-tailed deer.

Based on current information about where ticks reside and their preferred climate, this paper explores the possibility of using publicly-accessible remote sensing and climate data to predict the relative numbers of Lyme disease cases by county. These data from satellites and weather stations are compared to annual cases of Lyme disease in counties in the northeastern U.S. The hypothesis being tested is that, by county, mean temperature and precipitation in the three months leading up to the peak Lyme disease months and mean leaf area index one month before the heaviest Lyme disease months begin are positively and significantly associated with Lyme disease case numbers. Such associations could help to predict where the most Lyme disease cases are likely to appear which can help public health professionals to better allocate prevention and monitoring efforts.

2. Methods

The analysis focuses on counties within states in the northeastern U.S. (Maine, New Hampshire, Vermont, New York, Massachusetts, Rhode Island, Connecticut, Pennsylvania, Washington, D.C., West Virginia, New Jersey, Delaware, and Maryland). Data were collected for the years 2000 to 2018.

Outcome variable data—annual Lyme disease cases by county—were gathered from the CDC website.[1]

The predictor variables are mean precipitation, mean temperature, and mean leaf area index (LAI) in the period leading up to the the heaviest months of the annual Lyme disease season (starting in May). Average precipitation and average temperature by county from February to April of each year were collected from the NOAA climate data mapping website.[6] Precipitation is measured in inches, and temperature is measured in degrees Fahrenheit.

Leaf area index data were collected as a way of measuring the amount of grass and woodland area available for ticks to live in. LAI was downloaded for April 1 of each year (2000-2018), and is an eight-day composite of the best available data. LAI is measured in m^2 of green leaf area per m^2 of ground area. These data came from NASA's Terra MODIS system.[7] Mean values for LAI by county were calculated using the ArcGIS python package ArcPy. The necessary county boundaries were downloaded from ESRI.[8]

All data were cleaned and put into a format that contained an id field with the county, state, and year as well as the related value field. Data cleaning was done using Python and R.

Counties with no cases and those with case numbers above the 95th percentile were omitted from analysis. For each of these counties and across all years, there were a total of n= 4208 observations, with each observation consisting of the number of cases in each county for a given year, as well as the mean LAI on April 1 of that year, the mean precipitation from February to April of that year, and the mean temperature from February to April of that year.

Multivariate regression was performed with cases of Lyme disease as the outcome. This regression analysis was performed using R version 3.6.1.

3. Results

Results of multivariate regression, n=4208

	$\boldsymbol{\beta}$	**SE**	$\boldsymbol{t}$**-Statistic**	$\boldsymbol{p}$
Intercept	-11.7602105	8.23622823	-1.427864	1.534054e-01
Mean Temperature	0.8078330	0.20598136	3.921874	8.926401e-05
Mean Precipitation	1.9957074	0.36491759	5.468926	4.790608e-08
Mean LAI	0.3963812	0.02581917	15.352204	8.501209e-52
Adj. R^2	0.09			

Table 1: Results of multivariate regression with mean temperature, mean precipitation, and mean LAI by county by year (as outlined in the methods section) as predictors and Lyme disease cases by county by year as the outcome (n=2408)

Table 1 shows the results of the multivariate regression analysis and the adjusted R^2 from an overall F-test. As seen in the table, each of the coefficients were shown to be significant at the $\alpha = 0.05$ level.

Among counties in the northeastern United States that reported at least one case but less than the 95th percentile of cases for each year from 2000 to 2018, the following associations were observed:

1. Among two counties with the same mean precipitation from February to April of that year and the same LAI on April 1 of that year, the county with a mean temperature one degree Fahrenheit higher for the months of February to April of that year would report, on average, 0.81 more cases that year than the county with the lower mean temperature.
2. Among two counties with the same mean temperature from February to April of that year and the same LAI on April 1 of that year, the county with a mean precipitation one inch higher for the months of February to April of that year would report, on average, 2.00 more cases that year than the county with the lower mean precipitation.
3. Among two counties with the same mean temperature and mean precipitation from February to April of that year, the county with a mean LAI of $1m^2$ higher of green leaf area per $1m^2$ of ground surface on April 1 of that year would report, on average, 0.40 more cases than the county with the lower mean precipitation.

4. Discussion

The three predictors, mean precipitation, mean temperature, and mean LAI were shown to be significantly associated with an increase in reported Lyme disease cases, by year by county, which supports the hypothesis outlined in the introduction section.

The low R^2 value suggests that additional variables may be related to the outcome and therefore need to be included in the model in order to more accurately predict the number of cases. If it is possible in the future to gather annual data on deer density, urban buffer zones, humidity, population, or rodent density, these are likely worth including in further analysis given our understanding of tick behavior and Lyme disease transmission.

The overall model, however, is useful for generally predicting the relative case intensities by county, which was the hope for this project.

5. Conclusions

Precipitation, temperature, and LAI data collected in the months leading up to the peak of Lyme disease season in counties in the northeastern United States are associated with the number of Lyme disease cases reported in those counties for that year. These metrics, therefore, may be useful for public health officials to monitor when considering how and where to allocate resources and preventative interventions, as well as where they may want to monitor patients more closely for Lyme disease cases.

While typically used by climate or agricultural researchers, publicly-accessible remote sensing and weather station data can be of great value to public health researchers as well. Pursuing the use of remote sensing data to predict disease or other health outcomes—especially in combination with other public health data sets—is a worthwhile endeavor.

Limitations

As mentioned in the discussion section, predictive modeling of Lyme disease cases would likely benefit from additional variables. However, many of the variables that past research has suggested are likely associated with tick behavior, and therefore Lyme disease spread, were not available on an annual basis. Collection and incorporation of such variables is recommended for future analysis of Lyme disease spread and case numbers by geography.

It is also important to consider that case reporting procedures differ by state and year. Consistent reporting of Lyme disease cases across each county and year is obviously highly desirable, but was not practical without drastically reducing the sample size. In future years, it may be possible to gather more consistent data and to reach out to individual city- and state-level health departments to get more detailed information on collection and case information.

Acknowledgments. Thank you to Dr. Shira Dunsiger for her guidance on regression methods and general help with the project, and to Professor Lynn Carlson for her help with the ArcGIS components of this project.

References

[1] CDC - Lyme Disease Home[link].
URL `https://www.cdc.gov/lyme/datasurveillance/index.html`

[2] E. R. Adrion, J. Aucott, K. W. Lemke, J. P. Weiner, Health care costs, utilization and patterns of care following lyme disease, PLOS ONE 10 (2) (2015) e0116767. doi:10.1371/journal.pone.0116767.

[3] CDC - Lyme Disease Home[link].
URL `https://www.cdc.gov/lyme/stats/graphs.html`

[4] Tramsition, CDC - Lyme Disease Home.
URL `https://www.cdc.gov/lyme/transmission/index.html`

[5] J. Süss, C. Klaus, F. Gerstengarbe, P. C. Werner, What makes ticks tick? climate change, ticks, and tick-borne diseases, Journal of Travel Medicine 15 (1) (2008) 39–45. doi:10.1111/j.1708-8305.2007.00176.x.

[6] NOAA[link].
URL `https://www.ncdc.noaa.gov/cag/county/mapping`

[7] Y. K. R. Myneni, Mod15a2h modis/terra leaf area index/fpar 8-day l4 global 500m sin grid v006doi:10.5067/MODIS/MOD15A2H.006.
URL `https://lpdaac.usgs.gov/products/mod15a2hv006/`

[8] ESRI, I. TomTom North America, U. C. Bureau, U. D. of Agriculture (USDA), N. A. S. S. (NASS), Usa counties.
URL `https://www.arcgis.com/home/item.html?id=a00d6b6149b34ed3b833e10fb72ef47b`