Predicting Lyme Disease Cases in the Northeastern United States Using Remote Sensing and Weather Station Data

by Jack Hester, advised by Shira Dunsiger, Lynn Carlson, and Tyler Wray

BROWN
School of Public Health

Introduction

Lyme disease affects hundreds of thousands of Americans each year. While treatable, Lyme disease can cause severe illness and have lasting effects if not caught early. In addition to the health effects, there are hundreds of millions of dollars in associated annual costs. Based on current information about where ticks reside and their preferred climate, this project 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 incidence. 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.

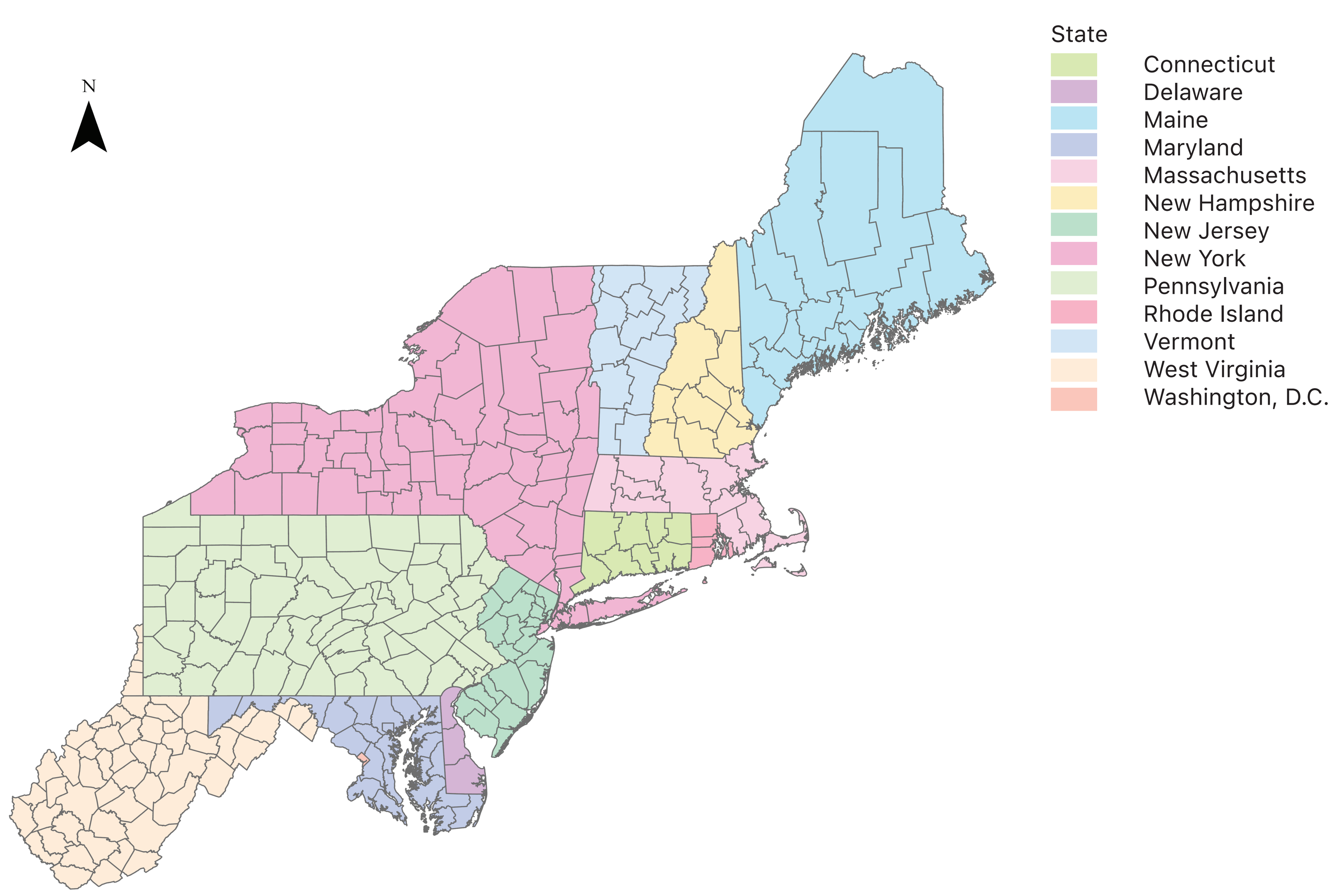
Methods

Lyme disease cases per thousand by county was the outcome of interest. The predictors were: 1) the mean precipitation (inches) by county for the three months leading up to May of that year (May is the beginning of peak Lyme disease season), 2) the mean temperature (°F) by county for those same three months, and 3) the mean leaf area index (LAI) (m² of green leaf area per m² of ground surface area) for the county on April 1 of that year. Analysis focused on states in the northeastern U.S. (see "Study Area" below). Data were collected for 2000-2018.

Lyme disease case data were downloaded from the CDC website,[1], precipitation and temperature data were downloaded from NOAA,[2] LAI data were downloaded from NASA's Terra MODIS system,[3] county boundaries were downloaded from ESRI,[4] and population estimates were downloaded from the national cancer institute.[5]

Data cleaning was performed in R and Python. Regression analysis (including assumption checking and overall model analysis) was performed in R 4.0.2. Mapping and LAI raster calculations were performed in ArcGIS Pro. Observations where the number of Lyme diseases cases was zero were omitted from analysis and from the maps as it was likely most of these observations were erroneous. The result was a multivariable regression model with n=4681 observations.

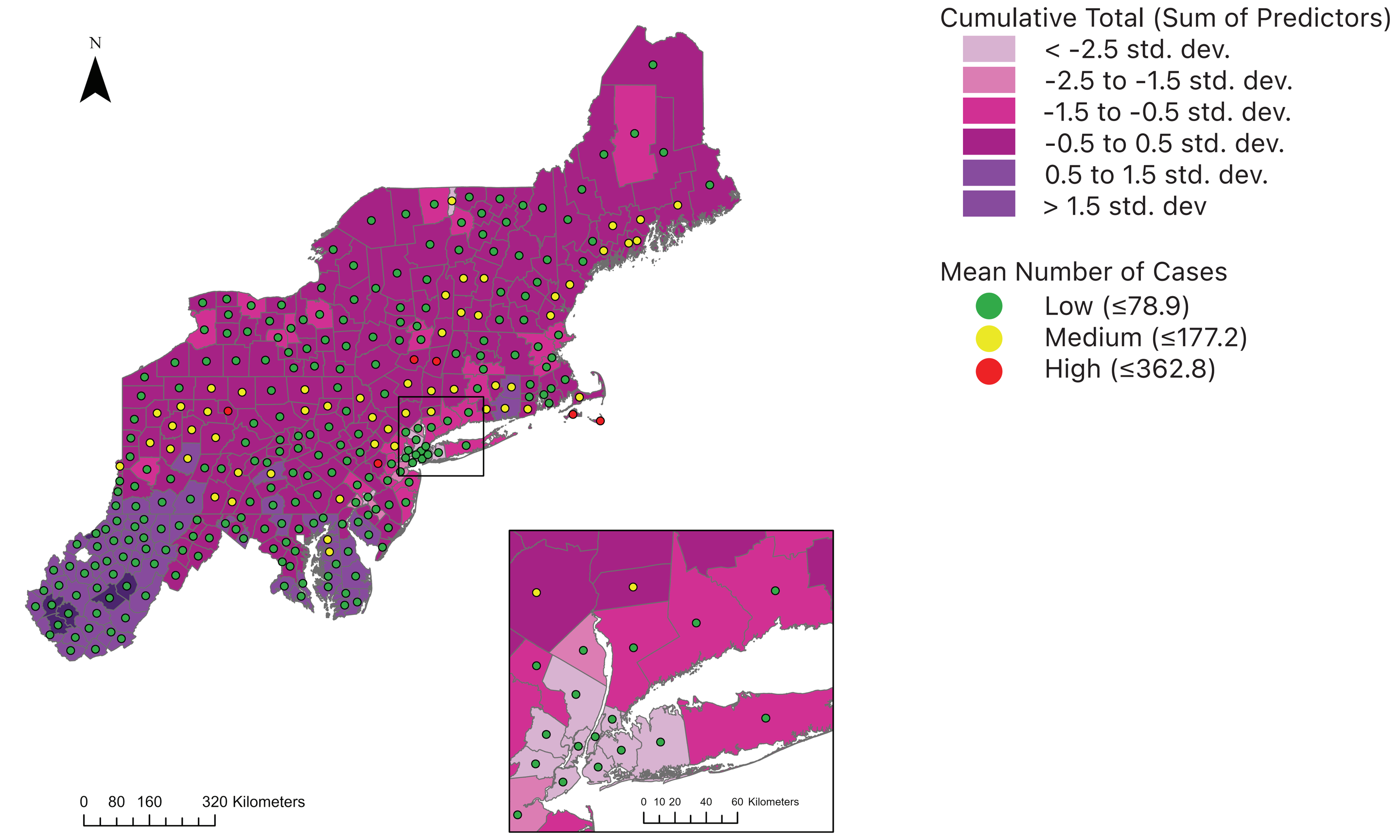
Study Area



Acknowledgements

A special thank you to Professor Lynn Carlson for all of her help in learning GIS and guidance throughout the project, and to Dr. Shira Dunsiger for her assistance with regression analysis. Thank you also to NASA, NOAA, NCI, and the CDC for making their data available.

Log Cases per Thousand vs. Weighted Cumulative Mean Score of All Predictors by County



The cumulative weighted score was created as a proxy for visualizing the regions with the highest likelihood of having an increase in the number of Lyme disease cases. The values are the result of putting the mean temperature, precipitation, and LAI of all years by county into the multivariable regression equation that was developed, without the intercept, to create a "weighted" score. The equation can be represented as *cumulative total=0.019*mean temp+0.043*mean precip-0.006*mean LAI*

Brief Results and Conclusions

All predictors were found to be significant at the $\alpha = 0.05$ level. For temperature, $\beta=0.019$, $SE=0.004$, and $p<0.001$. For precipitation, $\beta=0.043$, $SE=0.007$, and $p<0.001$. For LAI, $\beta=-0.006$, $SE=0.001$, and $p<0.001$. The model was found to predict Lyme disease cases per thousand, $F(3, 4734)=54.047$ and $p<0.001$, but the model only explained 3.3% of the variance, suggesting other variables need to be included to predict actual case numbers.

First, among two counties with the same mean precipitation from February to April of that year and the same mean LAI on April 1 in a given year, **the expected cases per thousand were 1.92% higher given a a one-degree Fahrenheit increase in temperature.**

Second, mong two counties with the same mean temperature from February to April of a given year and the same mean LAI on April 1 of that year, **the expected cases per thousand were 4.39% higher given a one-inch increase in precipitation.**

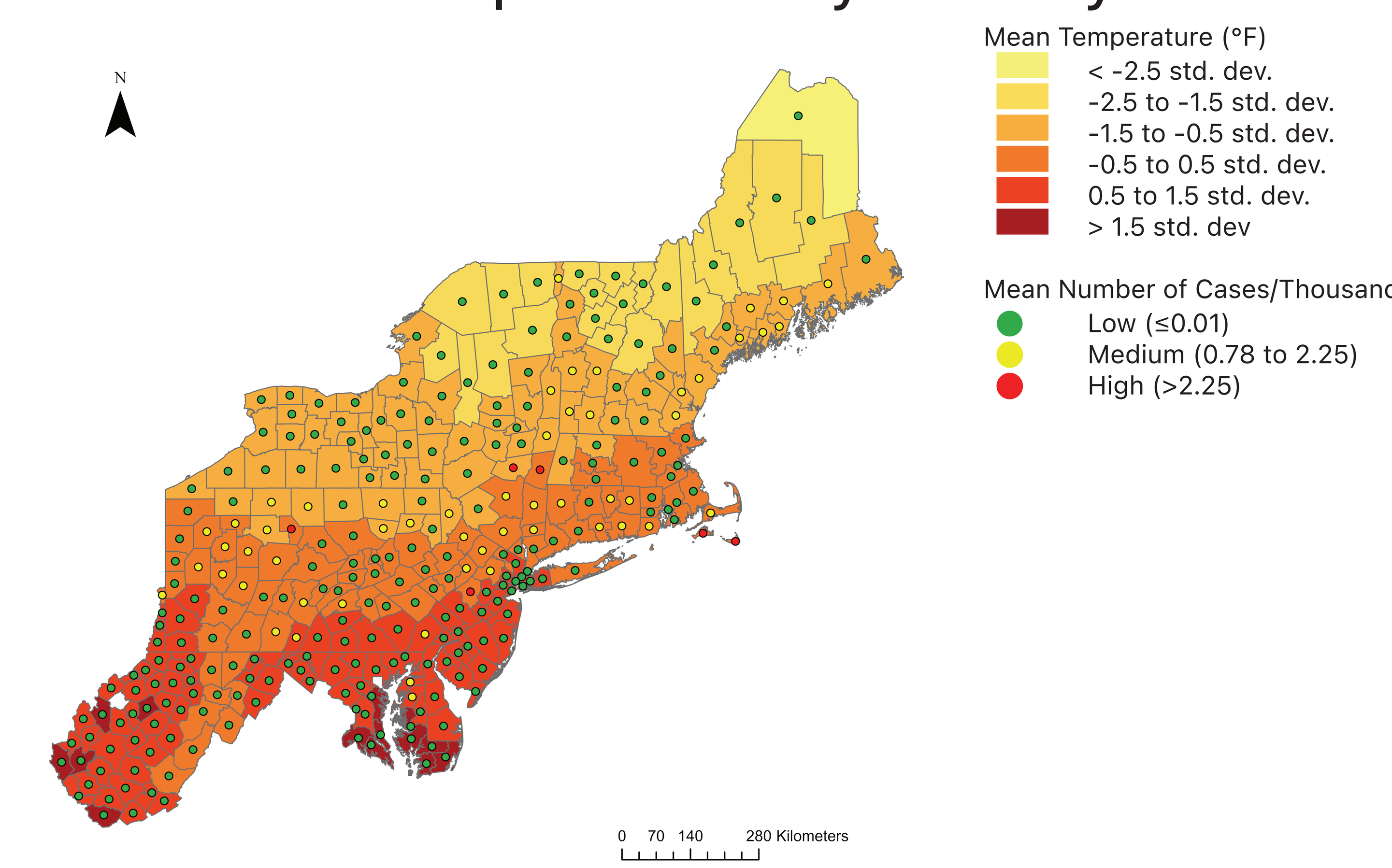
Third, among two counties with the same mean temperature and precipitation from February to April of a given year, **the expected cases per thousand were 0.60% lower given an increase of one square meter of green leaf coverage per square meter of ground area** (a one unit increase in LAI).

Based on these results, remote sensing data from satellites and weather stations seem to be a valuable source of information for predicting relative levels of Lyme disease. The association between LAI and cases is surprising, but may be explained by cases being reported in county of residence rather than where infection occured. More variables should be explored to facilitate more accurate predictions of cases per thousand, but this model could be useful when predicting relative case incidence for the upcoming Lyme disease season in the context of the northeastern U.S. Remote sensing data sources, especially in conjunction with other public health data, should be further explored as a way to inform disease modeling.

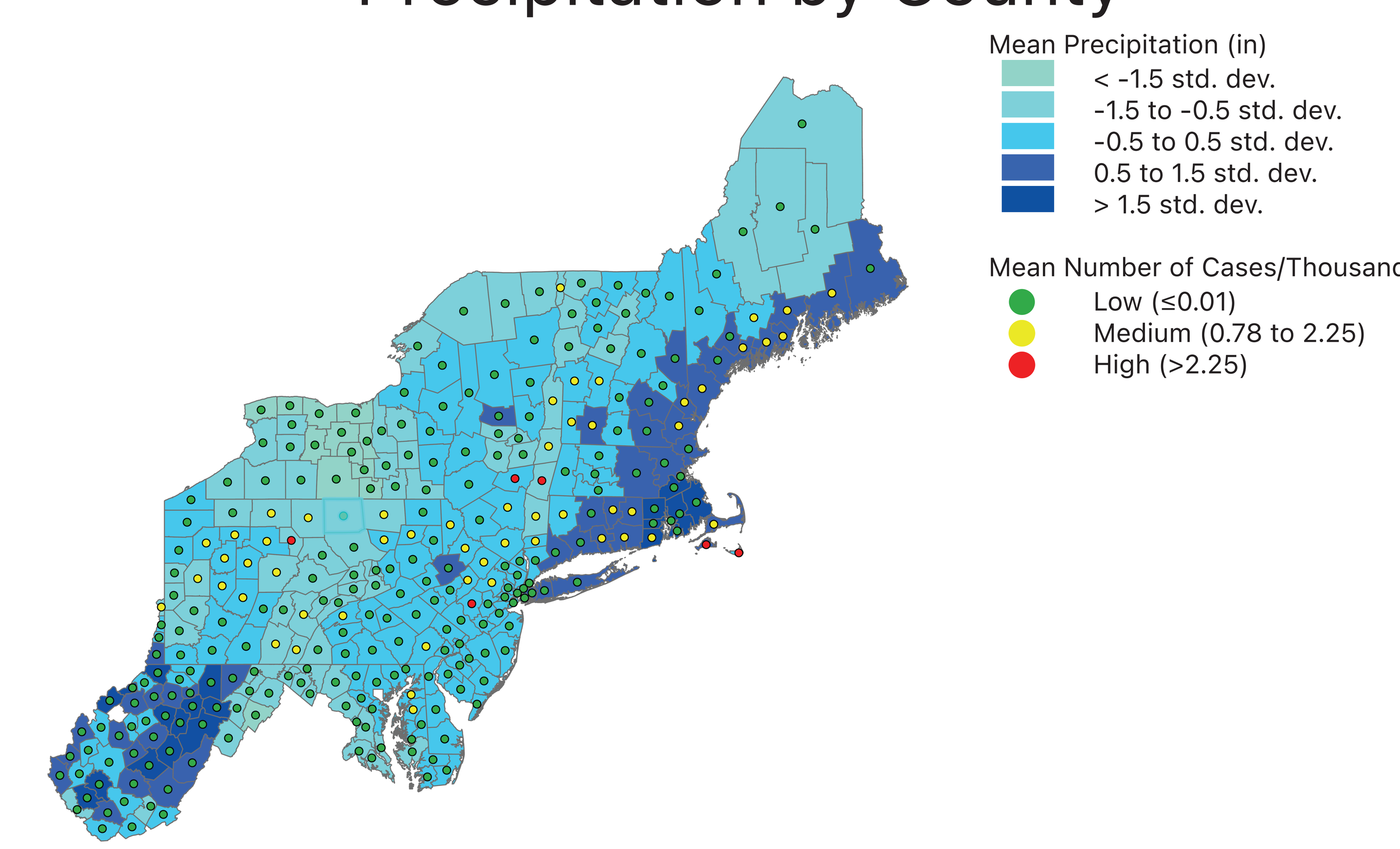
References

[1] Data and Surveillance [WWW Document], 2019. CDC - Lyme Disease Home. URL https://www.cdc.gov/lyme/datasurveillance/index.html
[2] NOAA National Centers for Environmental information, Climate at a Glance: County Mapping [WWW Document], 2020. NOAA.
[3] R. Myneni, Y.K., 2015. MOD15A2H MODIS/Terra Leaf Area Index/FPAR 8-Day L4 Global 500m SIN Grid V006. https://doi.org/10.5067/MODIS/MOD15A2H.006
[4] ESRI, TomTom North America, Inc., U.S. Census Bureau, U.S. Department of Agriculture (USDA), National Agricultural Statistics Service (NASS), n.d. USA Counties.
[5] Download U.S. Population Data - 1969-2019. National Institutes of Health | National Cancer Institute https://seer.cancer.gov/popdata/download.html#single (2021).

Cases per Thousand vs. Mean Temperature by County



Cases per Thousand vs. Mean Precipitation by County



Cases per Thousand vs. Mean LAI by County