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

Jack Hester

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 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.

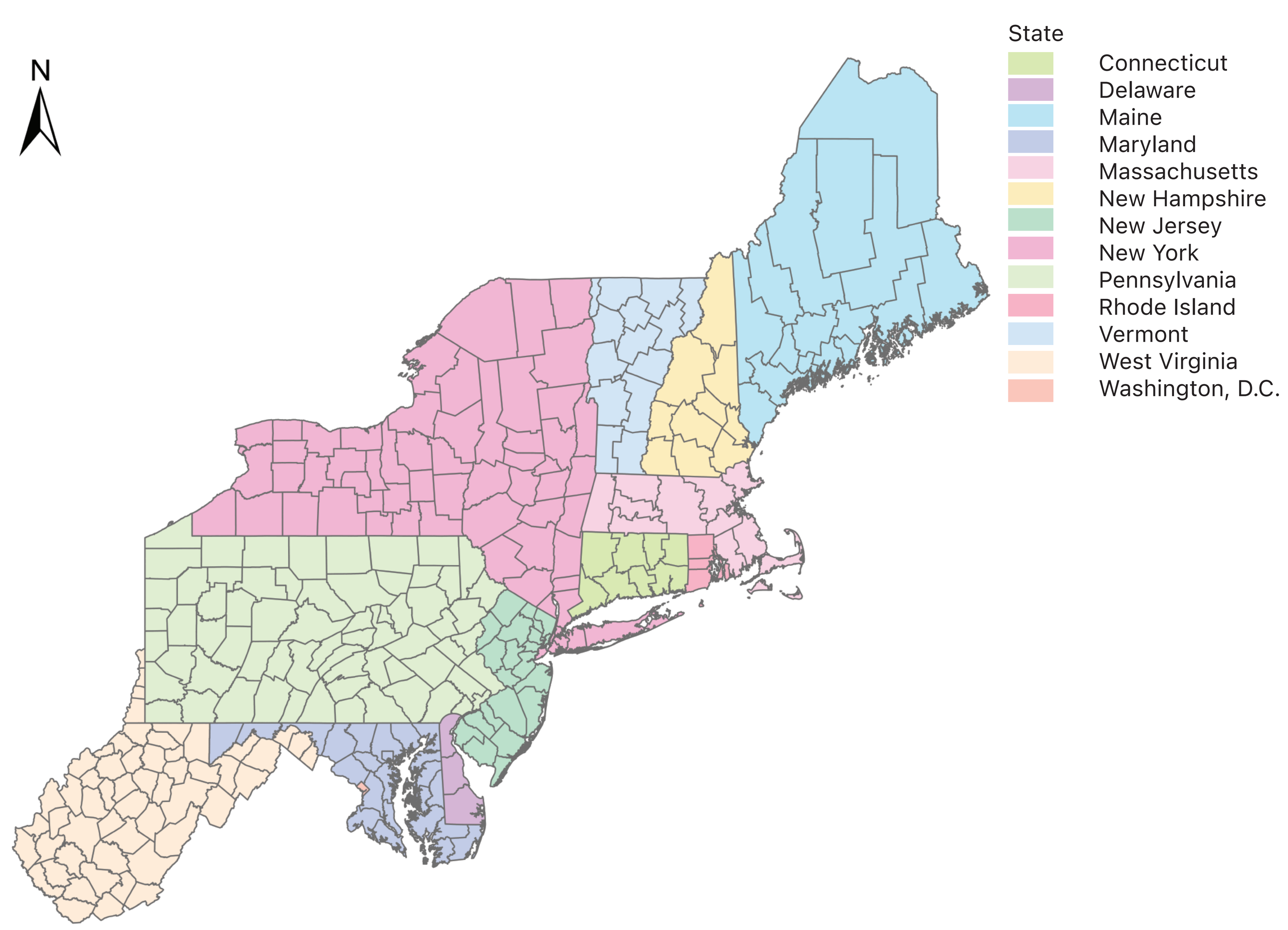
Methods

Lyme disease cases by county was the outcome of interest. The predictors were: 1) the mean precipitation (inches) by county for the three months leading up May of that year (May is the beginning of peak Lyme disease season), 2) the mean temperature (°F) by county for the same three months, and 3) the mean leaf area index (LAI) (m^2 of green leaf area per m^2 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] and county boundaries were downloaded from ESRI.[4]

Data cleaning was performed in R and Python. Regression analysis (including assumption checking and overall model analysis) was performed in R 3.6.1. Mapping and LAI raster calculations were performed in ArcGIS Pro. Regression analysis with n=4452 excluded time points where cases were zero or greater than 95% of the maximum number of cases. Mapping included all values.

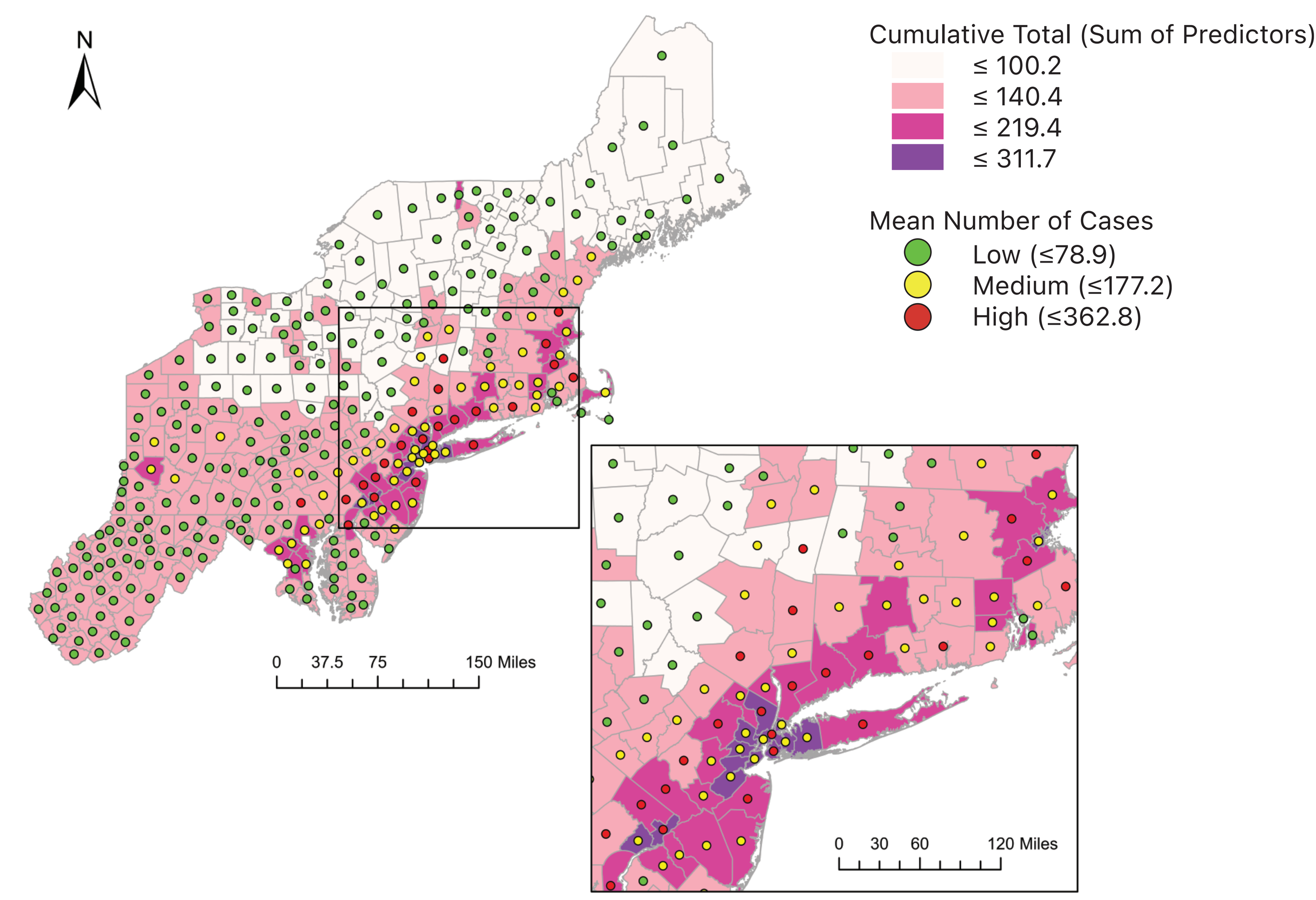
Study Area



Acknowledgements

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, and the CDC for making their data available.

Cases vs. Cumulative Mean Score by County for All Predictors Across All Years



The cumulative total was created as a proxy for visualizing the regions with the highest values for all of the predictors. It is a simple sum of the mean value, across all years 2000-2018, for each predictor

Brief Results and Conclusions

All predictors were found to be significant at the $\alpha = 0.05$ level. For precipitation, $\beta = 2.34$ and $p<.0001$. For temperature, $\beta=0.79$ and $p<.001$. For LAI, $\beta=0.41$ and $p<.0001$. The model was found to predict Lyme disease cases, $F(4, 4448) = 143.51$ and $p<.0001$, but the model only explained 8.76% of the variance, suggesting other variables need to be included to predict actual case numbers.

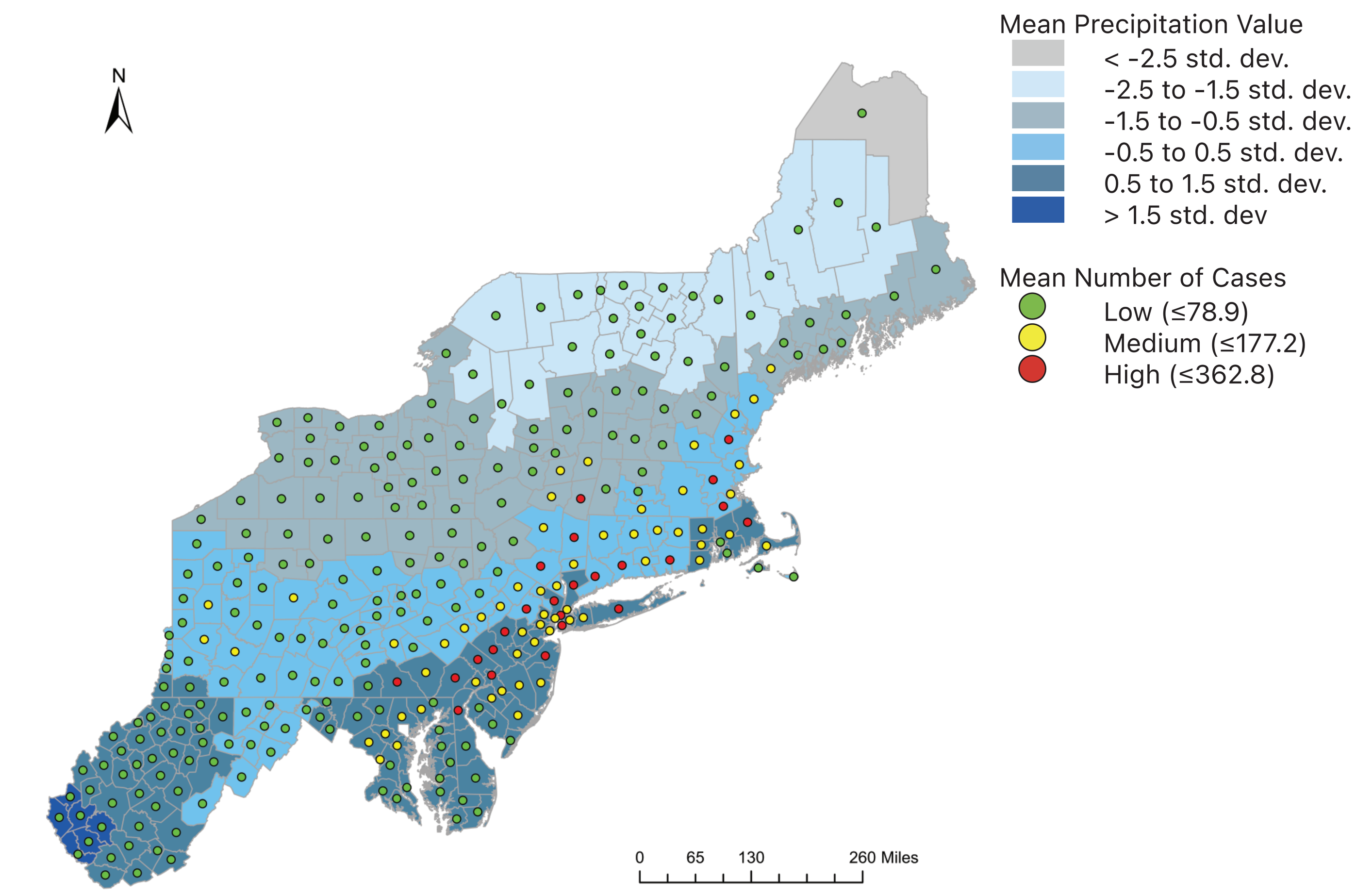
Among two counties with the same mean temperature from February to April of that year and the same LAI on April 1 of that year, the county with a mean precipitation one inch higher for the months of February to April of that year would report, on average, 2.34 more cases that year than the county with the lower mean precipitation. 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, 0.79 more cases that year than the county with the lower mean temperature. 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, 0.41 more cases than the county with the lower mean precipitation.

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. More variables should be explored to facilitate more accurate predictions of case numbers, 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.

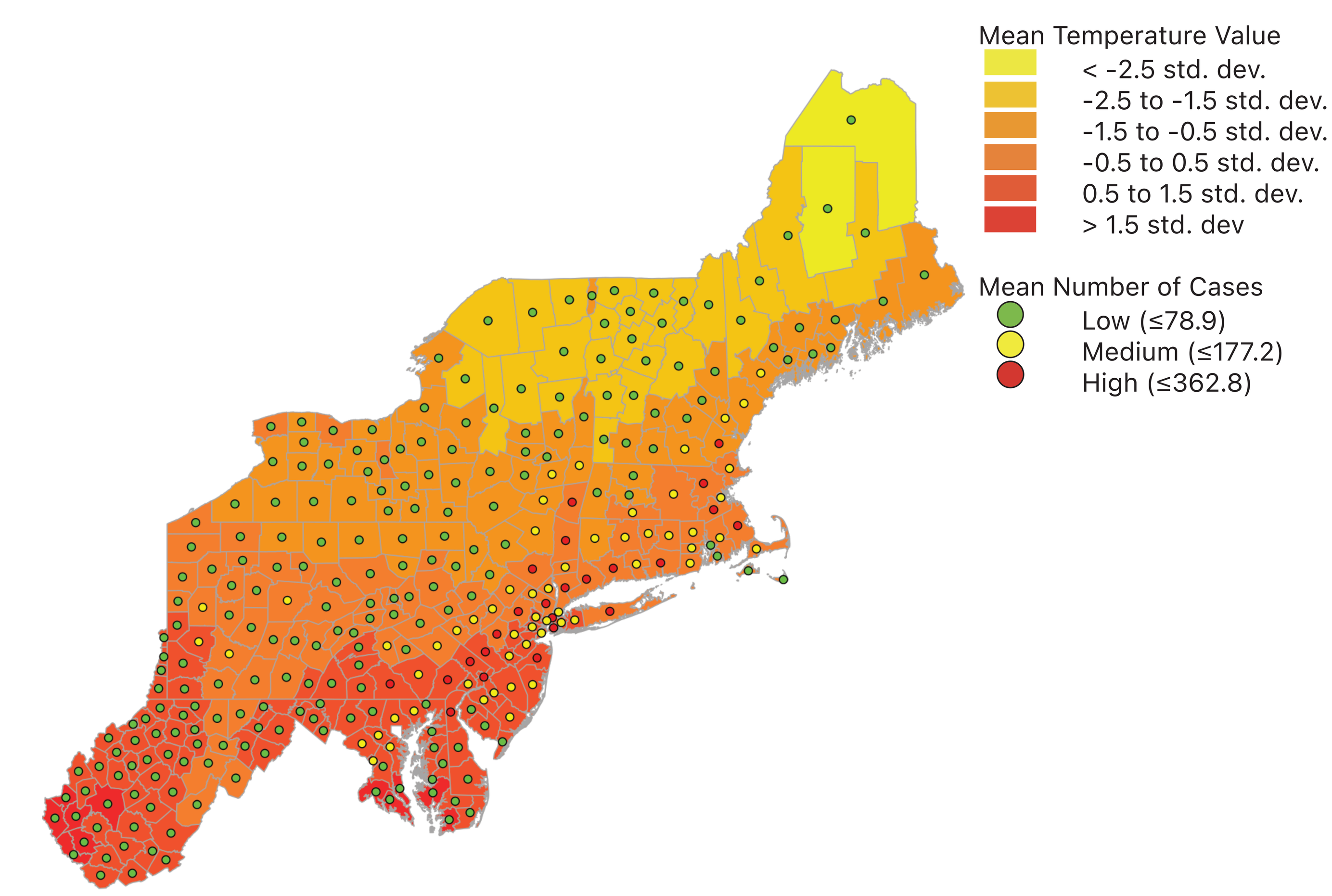
References

[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

Cases vs. Mean Precipitation by County



Cases vs. Mean Temperature by County



Cases vs. Mean LAI by County