Analysis of Lyme Disease Infection in the United States

As submitted by

Diwash Shrestha (12203784)

Daniel Eduardo Duran (12301739)

Email: diwash.shrestha@stud.th-deg.de Email: daniel.duran-vega@stud.th-deg.de

1. Introduction

In December 1975 two researchers led a surveillance study to investigate the cause of a sudden outbreak of rheumatoid arthritis in and around a town named Lyme. The investigation which consisted of physical and blood examinations discovered that 25% of patients presented a rash on their skin, furthermore, this mysterious arthritis emerged with interesting geographical and temporal patterns, as most patients lived in proximity to towns. All this information led the authors to discover a tick-transmitted infection called Lyme disease (Daniel Jernigan, 2023; Elbaum-Garfinkle, 2011). In this study, we will use data from Project Tycho to gain a deeper understanding of the Lyme Disease. With the help of this dataset, we will review the temporal and geographical patterns of the disease and the implications of our findings for public health prevention and control efforts.

2. Problem definition

In this work we will attempt to analyze the patterns of Lyme Disease. Specifically, we aim to analyze a data set from Project Tycho containing information on different diseases with their corresponding number of infections and death from 1888 to 2014. One challenge in this task is that the data is not evenly distributed among all diseases and may have certain limitations, such as underreporting bias or temporal bias due to the database not containing all reported years for Lyme Disease. Despite this, we expect this work to gain a deeper understanding of the factors that influence Lyme Disease and its behaviour in the United States.

3. Objectives

- Identify the temporal and geographical patterns of Lyme disease in the United States
- Evaluate the implications of these patterns for public health prevention and control efforts.

4. Methods

For the purposes of this analysis, we used the information provided by Project Tycho containing data on different diseases with their corresponding number of infections and death from 1888 to 2014. From the dataset, we selected Lyme Disease for this analysis as its different transmission vector from other diseases could expand our understanding of the patterns of tick-transmitted diseases. To obtain information on the temporal and geographical patterns of this disease we first identify the range in which Lyme Disease is found in the dataset and select the range of our assessment which for this work will include all available data from 2006 to 2014. Once our range was selected. We use advanced data visualization techniques such as the ones included in the libraries ggplot2, dygraphs, dygraphs, etc in R. With the help of this packages, we created bar charts, line charts and maps. Bar charts are a simple but effective way to visualize the distribution of data. We used bar charts to visualize the distribution of Lyme disease cases by state. Additionally, we used line plots as they are a good way to visualize the relationship between two variables such as cases and time or temperature, cases and time. These graphs will help us visualize the relationship between the number of Lyme disease cases and seasonality/temperature patterns. Finally, we will use Maps visualizations to observe the spatial distribution of data, which for this work will help us analyze the distribution of Lyme disease cases in the United States.

5. Analysis Protocol

The analysis was divided into subsections. The working steps are introduced below.

Before to start the analysis the required libraries were loaded

5.1 Data Loading and Cleanup

```
# Import Data
tycho_db <- fread("data/ProjectTycho_Level2_v1.1.0_0/ProjectTycho_Level2_v1.1.0.csv")
```

```
tycho_db %>% group_by(disease)%>%
  summarise(n = n())
```

```
## # A tibble: 50 x 2
##    disease                          n
##    <chr>                        <int>
##  1 ANTHRAX                       7051
##  2 BABESIOSIS                      36
##  3 BOTULISM                        10
##  4 BRUCELLOSIS [UNDULANT FEVER] 14970
##  5 CHICKENPOX [VARICELLA]       75198
##  6 CHLAMYDIA                    16837
##  7 CHOLERA                        125
##  8 COCCIDIOIDOMYCOSIS            1069
##  9 CRYPTOSPORIDIOSIS             7590
## 10 DENGUE                         621
## # i 40 more rows
```

Table 1. Disease summary of the dataset.

Data loading

Since we are analysing the lyme disease, we will filter data which contains the lyme disease.

```
# Filter the Lyme disease data
lyme_db <-tycho_db %>% filter(disease == "LYME DISEASE")
```

- **Data Cleanup**

As the data is from USA only and we are focused on the lyme disease. We will remove the country, disease and url column and keep other relevant information.

```
lyme_db <- lyme_db %>% select(-c('country','disease','url'))
```

- **Number of Cases and Deaths**

```
lyme_db %>% group_by(event)%>%
  summarise(sum(number))
```

```
## # A tibble: 1 x 2
##   event `sum(number)`
##   <chr>         <int>
## 1 CASES          76952
```

Table 2. Total cases reported for Lyme Disease in the dataset.

In total there were 76952 cases and no death cases

6. Result

In this section the overall trend of the cases occured in the USA was studied.

6.1 Yearly cases trend

To initialize our analysis of the Lyme Disease cases occured in the Unites States, we first look into the amount of cases occurred in each year

```
# break the epi_week column which has the year and week data combined
yr_wk_case_db <- lyme_db %>% group_by(epi_week)%>%
  summarise(cases = sum(number)) %>% separate(epi_week, into = c('year', 'week'), sep = -2, convert = T
yr_wk_case_db
```

```
## # A tibble: 447 x 3
##     year  week cases
##    <int> <int> <int>
##  1  2006     1    14
##  2  2006     2    17
##  3  2006     3    22
##  4  2006     4    22
##  5  2006     5    35
##  6  2006     6    33
##  7  2006     7    31
##  8  2006     8    23
##  9  2006     9    39
## 10  2006    10   144
## # i 437 more rows
```

Table 3. Total cases per week of the year.

The resulted data frame has the data in the range of January 2006 to August 2014. As a consequence, We will only take data from 2006 to 2013 to have the uniformity for the yearly trend analysis.

```
# group by yearly and find the cases on yearly basis
yr_case <- yr_wk_case_db%>%group_by(year)%>%
  summarise(cases = sum(cases)) %>%
  filter(year != 2014)
```

As we want to make the visualization uniform in our anlysis. A theme for the visualization is created for this analysis.

```
# theme for visualization
my_theme <-   theme_minimal(base_size = 24)+theme(plot.title = element_text(hjust = 0.5,face = "bold"),
        plot.subtitle = element_text(hjust = 0.5,face = "italic"),legend.position="top")
```

- **Bar Plot**

```
p <- ggplot(yr_case,aes(year,cases))+
  geom_col(width=0.5, fill="royalblue",color="royalblue")+
  geom_label(aes(label=cases),size=6)+
  scale_x_continuous(breaks=c(2006,2007,2008,2009,2010,2011,2012,2013,2014))+
  labs(title = "No. of Lyme Disease cases in USA",
       subtitle = "2006 to 2013",
       x = "Year",
       y = "Cases")+
  my_theme

p
```

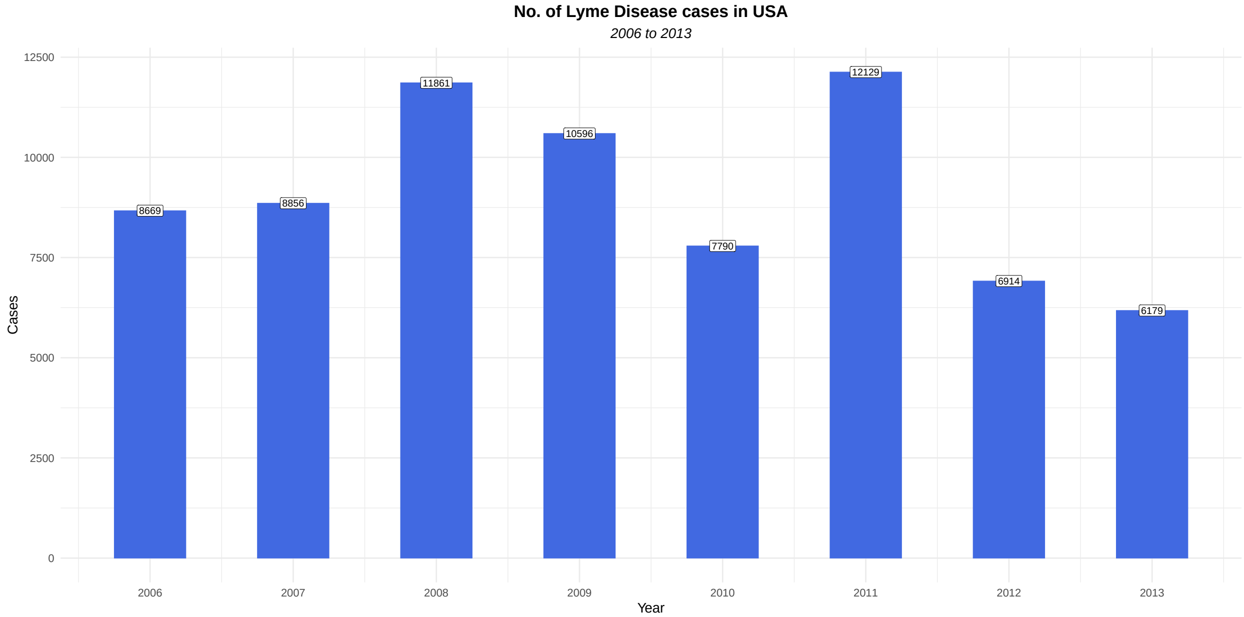

Fig 1. Bar-plot of the total cases vs year.

The yearly graph shows that the number of Lyme disease cases in the United States has a high variance from 2006 until 2011, finding the highest number of cases in 2011, with over 1200 cases. Despite this, the number of cases then decreased consistently from 2012 onwards finding them minimum in 2013 with 6156.

- **Monthly Trends**

Given the high disparity in our data we decided to take the monthly pattern into account. To achieve this we are used the to_date columns to calculate the cases found till that date.

```
year_month_cases <- lyme_db %>% group_by(to_date)%>%summarise(cases = sum(number))%>%separate(to_date, 
  group_by(year_month)%>%
  summarise(cases = sum(cases))

year_month_cases$year_month <- zoo::as.yearmon(year_month_cases$year_month, format="%Y-%m")
```

- **Line Chart**

For the monthly trend we decided to use a line chart given that this graph is better suited to provide us with the trend analysis for the amount of monthly data given by the project.

```
p <- ggplot(year_month_cases, aes(as.Date(year_month), cases)) +
  geom_line(color="royalblue",size=1)+
    geom_point(color="red",size=3)+
  scale_x_date(date_breaks = "6 month",date_labels = "%Y-%m")+
  labs(title = "Timeline of the cases in each month",
       subtitle = "From 2006-January to 2014-August",
       x="Year and Month", y="No of Cases")+
  my_theme
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
p
```

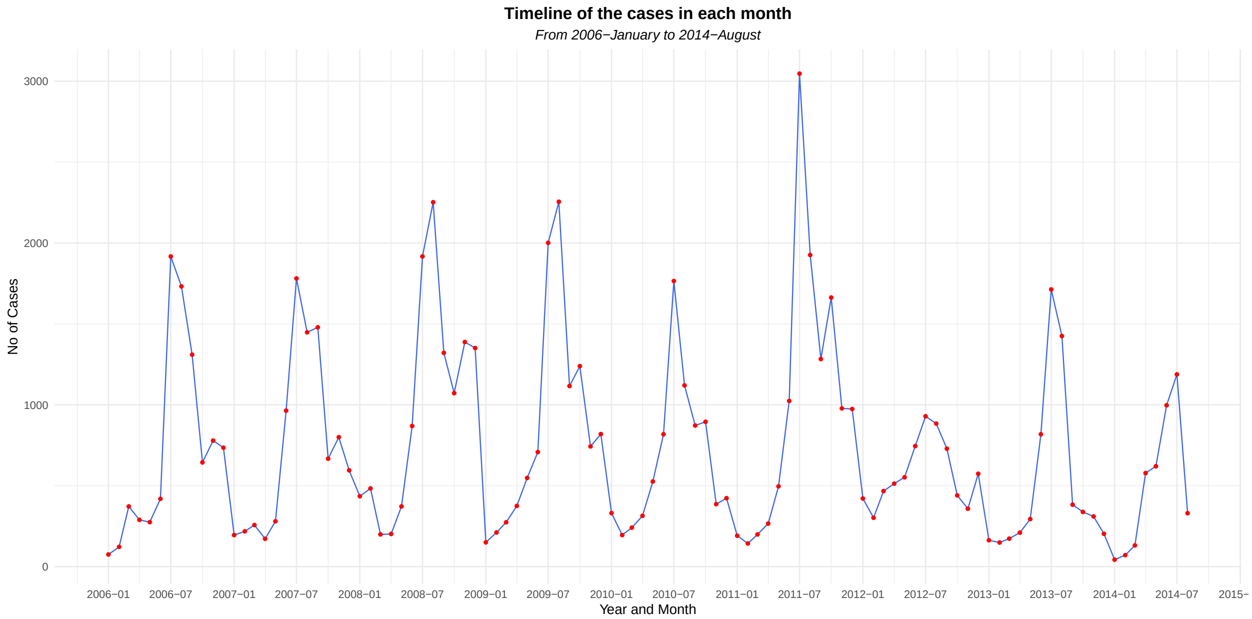

Fig 2. Graph of Lyme Disease cases along the decided timeline.

The results of our line plot showed that Lyme Disease has a seasonality pattern, defined by a sharp increase of cases starting from March until August followed by an acute decline of cases from October to February. This pattern is consistently repeated every year.

6.2 Geographical analysis

In this section the occurrence of lyme disease will be discussed based on the state and region of the United States.

- **Top 10 states with highest lyme disease cases**

```
total_case_state <- lyme_db %>% group_by(state)%>%
  summarise(cases = sum(number))%>%arrange(-cases)
```

```
# top 10 state with most cases
top_10_db <- lyme_db %>% group_by(state)%>%
  summarise(cases = sum(number))%>%arrange(-cases) %>%head(10)
top_10_db$state_name <-c("New York","Pennsylvania","Maryland","Virginia","Connecticut","New Jersey","Mi
top_10_db
```

```
## # A tibble: 10 x 3
##    state cases state_name
##    <chr> <int> <chr>
##  1 NY    28855 New York
##  2 PA    20121 Pennsylvania
##   3 MD      5211 Maryland
##   4 VA      3564 Virginia
##   5 CT      3393 Connecticut
##   6 NJ      3088 New Jersey
##   7 MN      2575 Minnesota
##   8 ME      2321 Maine
##   9 DE      1319 Delware
##  10 FL      1119 Florida
```

Table 4. Top 10 states from cases of Lyme Disease.

```
# Top 10 state with cases
p <- ggplot(data = top_10_db)+
  geom_col(aes(reorder(state_name,-cases),cases),width=0.5, fill = "royalblue")+
  geom_label(aes(label=cases,reorder(state_name,-cases),cases),size=6)+
  labs(title = "Top 10 States with highest cases",
       x = "States",
       y = "Cases")+
  my_theme
p
```

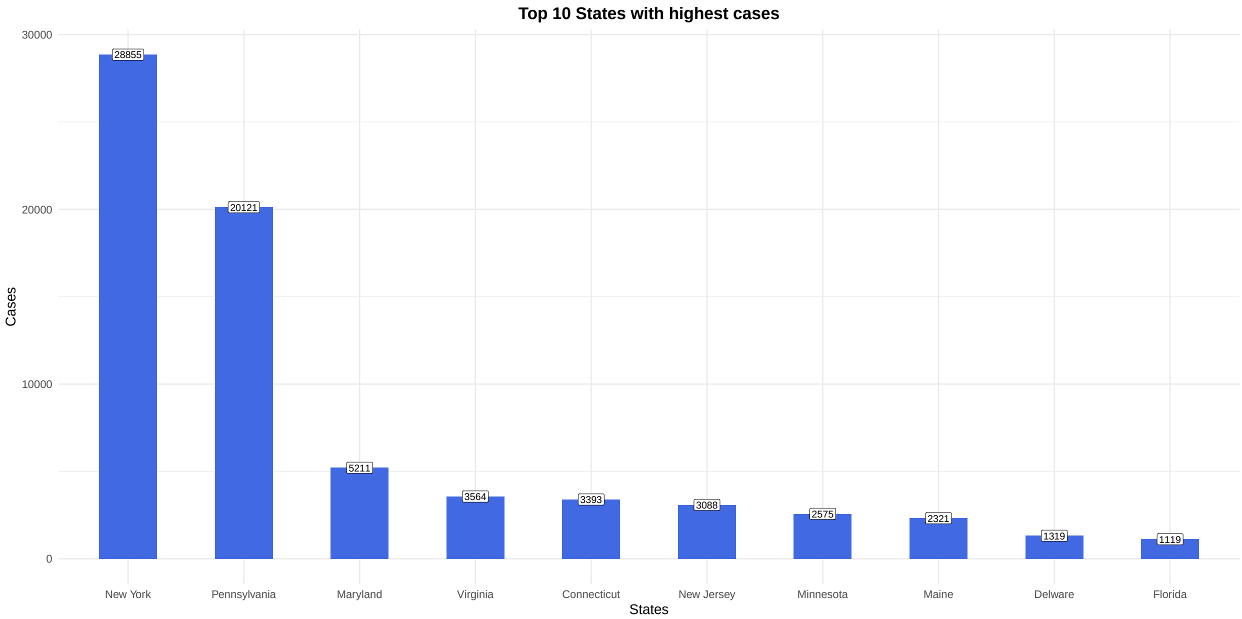

Fig 3. Bar-plot of the total cases vs state.

- **Choropleth Map**

The Choropleth Map visualization technique make it easier to compare the data in a given map, as it colors the area with larger number with darker color and smaller one with lighter color. Hence, we used it to increase our understatement of geographical patterns in the USA.

```
plot_usmap(data = total_case_state, values = "cases", color = "black") +
  scale_fill_continuous(
    low = "white", high = "red", name = "Cases", label = scales::comma
  ) +labs(title = "Total Lyme Disease Cases  by State",
       subtitle = "From 2006-January to 2014-August",
       x="", y="")+
  theme(plot.title = element_text(hjust = 0.5,face = "bold",size = 28),
        plot.subtitle = element_text(hjust = 0.5,face = "italic",size = 16),legend.position = "right")
```

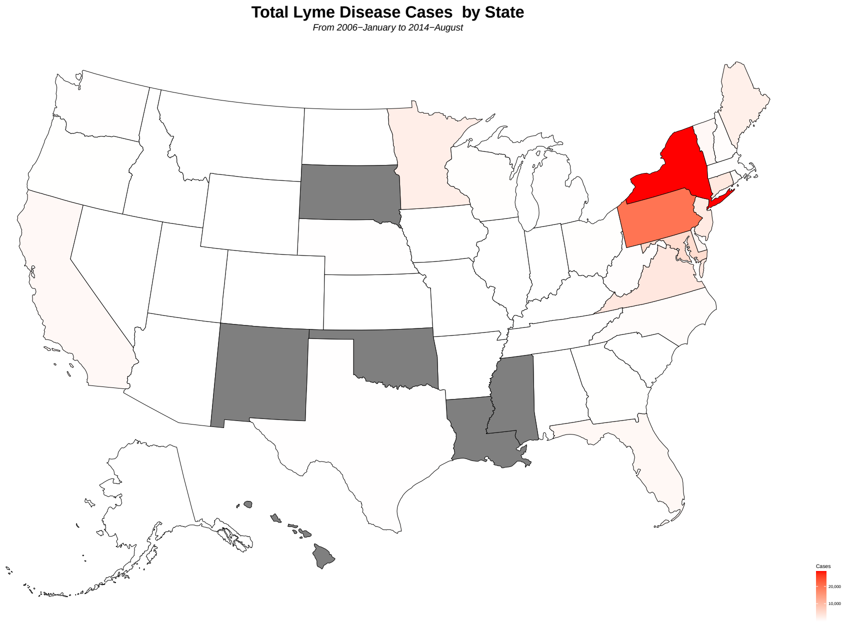
Total Lyme Disease Cases by State
From 2006−January to 2014−August
Cases
20,000
10,000

Fig 4. Choropleth Map of the united states for Lyme Disease cases.

Our spatial analysis shows that Lyme Disease is concentrated in the northeast of the united states, especially in the states of New York and Pennsylvania, which in conjunction contain the more than 50% of all the cases.

6.3 Trends in Northeast and Other State

Looking into the choropleth map we wanted to look further as most of the cases seems to appears in northeastern state. Hence, we compared the northeast region with other remaining state.

```
northeast_st <- c('CT','ME','MA','NH' ,'NJ','NY','PA','RI','VT')
```

```
north_east_other_year_trend <- lyme_db %>% group_by(to_date,state)%>%summarise(cases = sum(number))%>%se
  mutate(region = case_when(
  state %in% northeast_st ~ "northeastern",
  .default = "other"
))%>%
  group_by(year_month,region)%>%
  summarise(cases = sum(cases))
```

```
## `summarise()` has grouped output by 'to_date'. You can override using the
## `.groups` argument.
## `summarise()` has grouped output by 'year_month'. You can override using the
## `.groups` argument.
```

```
north_east_other_year_trend$year_month <- zoo::as.yearmon(north_east_other_year_trend$year_month, forma
```

```
north_east_other_year_trend
```

```
## # A tibble: 208 x 3
## # Groups:   year_month [104]
##    year_month region       cases
##    <yearmon>  <chr>        <int>
##  1 Jan 2006   northeastern    40
##  2 Jan 2006   other           35
##  3 Feb 2006   northeastern    78
##  4 Feb 2006   other           44
##  5 Mar 2006   northeastern   292
##  6 Mar 2006   other           80
##  7 Apr 2006   northeastern   260
##  8 Apr 2006   other           29
##  9 May 2006   northeastern   219
## 10 May 2006   other           56
## # i 198 more rows
```

Table 5. Northeastern cases against the rest of the states in the country.

```
p <- ggplot(north_east_other_year_trend, aes(as.Date(year_month), cases)) +
  geom_point(aes(color=region),size=3)+
  geom_line(aes(color=region),size=1)+
  scale_x_date(date_breaks = "6 month",date_labels = "%Y-%m")+
  labs(title = "Timeline of the cases in each month",
       subtitle = "From 2006-January to 2014-August",
       x="Year and Month", y="No of Cases")+
  my_theme
p
```

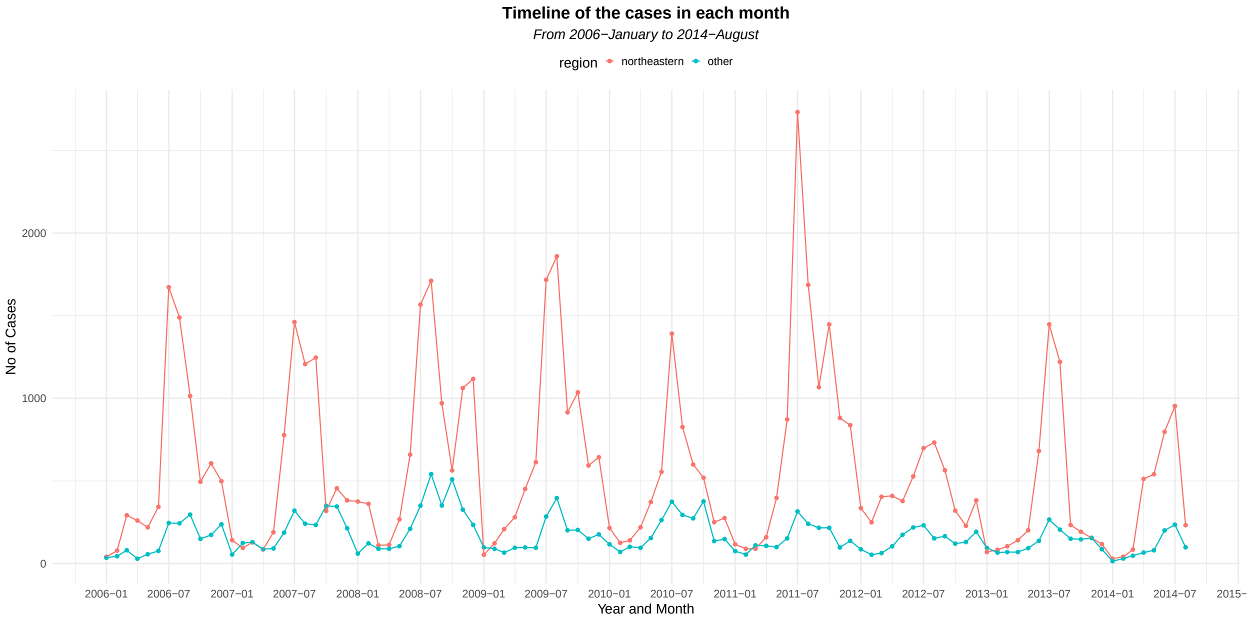

Fig 5. Temperature + total cases vs time.

6.4 Temperature and Lyme Cases

As we found the most of the cases are seen in the northeast state, you should remember we also notice the most effected states are New York and Pennsylvania. In this section we will focus on the temperature and cases of these states.

We extracted the data of the monthly average temperature of the Pennsylvania and new York to compare the cases and the temperature.

Temperature data was taken from (NOAA National Centers for Environmental Information, 2023).

```
  # load the temperature data
temp_db <- fread("data/NY_PN_temp_data.csv")
temp_db$date <- zoo::as.yearmon(temp_db$date, format="%Y-%m")
```

```
head(temp_db)
```

```
##        date    state value anamoly
## 1: Jan 2006 New York  30.1     8.5
## 2: Feb 2006 New York  24.7     1.2
## 3: Mar 2006 New York  31.7     0.1
## 4: Apr 2006 New York  45.7     1.8
## 5: May 2006 New York  55.7    -0.1
## 6: Jun 2006 New York  64.4     0.0
```

Table 6. Head of the average temperature from New York and Pennsylvania states.

This data frame has the temperature in Fahrenheit of the New York and Pennsylvania state. In order to have a better representation we decided to convert the temperature in Fahrenheit to Celsius scale.

```
temp_db <- temp_db %>%
  mutate(celcius_value = weathermetrics::fahrenheit.to.celsius(value))

temp_db
```

```
##          date        state value anamoly celcius_value
##   1: Jan 2006     New York  30.1     8.5         -1.06
##   2: Feb 2006     New York  24.7     1.2         -4.06
##   3: Mar 2006     New York  31.7     0.1         -0.17
##   4: Apr 2006     New York  45.7     1.8          7.61
##   5: May 2006     New York  55.7    -0.1         13.17
##  ---
## 212: Aug 2014 Pennsylvania  66.9    -2.6         19.39
## 213: Sep 2014 Pennsylvania  62.5    -0.3         16.94
## 214: Oct 2014 Pennsylvania  53.0     1.6         11.67
## 215: Nov 2014 Pennsylvania  36.7    -4.0          2.61
## 216: Dec 2014 Pennsylvania  34.0     2.3          1.11
```

Table 7. Data frame of the average temperature from New York and Pennsylvania states in Celsius.

```
# convert the fahrenheit anamoly  to celcius anamoly

temp_db <- temp_db %>%
  mutate(celcius_anamoly = (anamoly) *(5/9))
```

```
# extract the total cases of the New York and Pennsylvania on monthly basis from 2006 to 2014
ny_pa_cases <- lyme_db %>% group_by(to_date,state)%>%summarise(cases = sum(number))%>%separate(to_date,
  filter(state %in% c('NY','PA'))%>%
  group_by(year_month, state)%>%
  summarise(cases = sum(cases))
```

```
## `summarise()` has grouped output by 'to_date'. You can override using the
## `.groups` argument.
## `summarise()` has grouped output by 'year_month'. You can override using the
## `.groups` argument.
```

```
ny_pa_cases$year_month <- zoo::as.yearmon(ny_pa_cases$year_month, format="%Y-%m")
ny_pa_cases
```

```
## # A tibble: 204 x 3
## # Groups:   year_month [104]
##    year_month state cases
##    <yearmon>  <chr> <int>
##  1 Jan 2006   NY       16
##  2 Jan 2006   PA       22
##  3 Feb 2006   NY       55
##  4 Feb 2006   PA       15
##  5 Mar 2006   NY      259
##  6 Mar 2006   PA       17
##  7 Apr 2006   NY      240
##  8 Apr 2006   PA       14
##  9 May 2006   NY      159
## 10 May 2006   PA       35
## # i 194 more rows
```

```
# change the name of the year_month column to date and change the value in state column to full name
# use the two columns to combine the ny_pa_cases and temp_db dataframe

ny_pa_cases <- ny_pa_cases %>%
  rename(date = year_month)%>%
  mutate(state = case_when(state == 'NY' ~'New York',
                           .default = 'Pennsylvania'))

ny_pa_cases
```

```
## # A tibble: 204 x 3
## # Groups:   date [104]
##    date      state        cases
##    <yearmon> <chr>        <int>
##  1 Jan 2006  New York        16
##  2 Jan 2006  Pennsylvania    22
##  3 Feb 2006  New York        55
##  4 Feb 2006  Pennsylvania    15
##  5 Mar 2006  New York       259
##  6 Mar 2006  Pennsylvania    17
##  7 Apr 2006  New York       240
##  8 Apr 2006  Pennsylvania    14
##  9 May 2006  New York       159
## 10 May 2006  Pennsylvania    35
## # i 194 more rows
```

Table 8. Cases of New York state and Pennsilvanya sate per month and year.

```
# Join the two dataframe using the state and date as ID
ny_pa_temp_case_db <- inner_join(ny_pa_cases, temp_db,by = c( "date","state"))
ny_pa_temp_case_db
```

```
## # A tibble: 204 x 7
## # Groups:   date [104]
##    date      state        cases value anamoly celcius_value celcius_anamoly
##    <yearmon> <chr>        <int> <dbl>   <dbl>         <dbl>           <dbl>
##  1 Jan 2006  New York        16  30.1     8.5         -1.06          4.72
##  2 Jan 2006  Pennsylvania    22  35.2     8.6          1.78          4.78
##  3 Feb 2006  New York        55  24.7     1.2         -4.06          0.667
##  4 Feb 2006  Pennsylvania    15  29.8     0.9         -1.22          0.5
##  5 Mar 2006  New York       259  31.7     0.1         -0.17          0.0556
##  6 Mar 2006  Pennsylvania    17  37       0.2          2.78          0.111
##  7 Apr 2006  New York       240  45.7     1.8          7.61          1
##  8 Apr 2006  Pennsylvania    14  50.5     2.1         10.3           1.17
##  9 May 2006  New York       159  55.7    -0.1          13.2           -0.0556
## 10 May 2006  Pennsylvania    35  57.3    -1.3          14.1           -0.722
## # i 194 more rows
```

Table 9. Joint data frame of cases of New York state and Pennsilvanya sate per month and year with temperature.

```
coeff = 20
p <- ggplot(ny_pa_temp_case_db, aes(x = as.Date(date)) +
  geom_col(aes(y = cases,fill=state))+
  geom_point(aes(y = (celcius_value*coeff)))+
  geom_line(aes(y = (celcius_value*coeff)))+
  scale_x_date(date_breaks = "6 month",date_labels = "%Y-%m")+
  facet_wrap(~state,ncol=1)+
  scale_y_continuous(

    # Features of the first axis
    name = "No of Cases",

    # Add a second axis and specify its features
    sec.axis = sec_axis(~./coeff, name="Average Temperature in °C")
  )+
  labs(title = "Comparision of the monthly cases and temperature  in New York and Penssylvania",
      subtitle = "Data from 2006 to 2014",x="Date in Year-Month"
      )+
  my_theme+ theme(legend.position="")
p
```

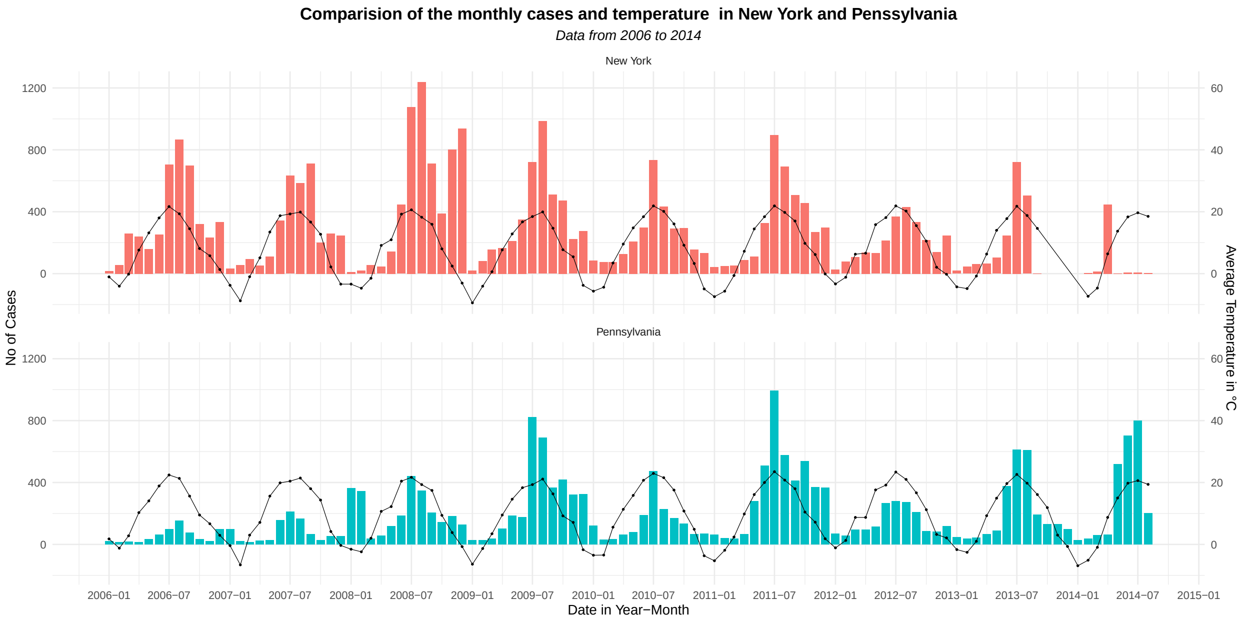

Fig 6. Temperature + total cases vs time

7. Discussion

In this study, we explored the correlations of climatic and geographical factors with Lyme cases at a country level in the United States. The results from our fig 3 and fig 4 analysis suggested that the geographical factors related to the state location between the United States has an incidence in the increase of cases of

Lyme disease. Laboratory studies had shown that ticks are highly vulnerable to desiccation and generally had high mortality in conditions with low humidity and high temperature(Dong et al., 2020). Thus, temperature and humidity may affect Lyme disease risk indirectly through the impacts on tick survivals and population dynamics. The high concentration of Lyme disease cases in the northeastern region of the United States can be partially attributed to its humid and hot climate during the summertime. This observation is supported by our results from Figures 2 and 5, which clearly demonstrate an increasing trend of infection with rising temperatures. Thus, the prevalence of favorable weather conditions for ticks in this region contributes to a higher number of infections. New York and Pennsylvania, located in the northeastern region, account for most reported cases, as evident in the figures 3 and 4. This is further exacerbated when we divide our temporality graph by states as can be seen in the figure 5, which clearly shows, the northeastern cases account for the majority of the reports. Some authors relate the number of cases from these states to reforestation of the forest, the increase in the deer population which carries the ticks that lead to Lyme Disease and climate change.(Ginsberg et al., 2021; VanAcker et al., 2019) We believe further studies must be carried out to explain this abnormality, as finding the main causes could led to environmental policies which could reduce dramatically the amount of infection of Lyme disease. We must admit that the project Tycho dataset falls short on information to help us find out policies to solve its causes. As only the amount of cases per county between 2006 and 2014 is insufficient to encounter a meaning public policy that could bring an end to the disease. Figure 1, which visualizes the yearly trend data, supports this claim.

8. Conclusion

Our results show that epidemiological research is important to explain or predict spatial patterns of observed communicable disease outcomes using a growing body of spatial data and tools, as well as spatial statistical methods including but not limited to the qualification of spatial mapping. We observed that Lyme Disease is affected by location and weather patterns, in which mainly New York and Pennsylvania are the main focus of infection of this disease and that it is mainly within the summer season where most of the infections are presented. Finally, it would be very informative for future research to identify why are these states significantly more affected by this disease and which state policies could help its reduction.

Literature

- Daniel Jernigan. (2023, January 20). Lyme Disease Trasmission. https://www.cdc.gov/lyme/transmission/index.html Accessed 28 Jun 2023.

- Dong, Y., Huang, Z., Zhang, Y., Wang, Y. X. G., & La, Y. (2020). Comparing the climatic and landscape risk factors for lyme disease cases in the upper midwest and Northeast United States. International Journal of Environmental Research and Public Health, 17(5). https://doi.org/10.3390/ijerph17051548

- Elbaum-Garfinkle, S. (2011). Close to home: a history of Yale and Lyme disease. The Yale Journal of Biology and Medicine, 84, 103–108. Ginsberg, H. S., Hickling, G. J., Burke, R. L., Ogden, N. H., Beati, L., LeBrun, R. A., Arsnoe, I. M., Gerhold, R., Han, S., Jackson, K., Maestas, L., Moody, T., Pang, G., Ross, B., Rulison, E. L., & Tsao, J. I. (2021). Why Lyme disease is common in the northern US, but rare in the south: The roles of host choice, host-seeking behavior, and tick density. PLoS Biology, 19(1). https://doi.org/10.1371/journal.pbio.3001066

- NOAA National Centers for Environmental Information. (2023, June 30). Climate at a Glance: Statewide Time Series. Https://Www.Ncei.Noaa.Gov/Access/Monitoring/Climate-at-a-Glance/Statewide/Time-Series. Accessed 30 Jun 2023.

- VanAcker, M. C., Little, E. A. H., Molaei, G., Bajwa, W. I., & Diuk-Wasser, M. A. (2019). Enhancement of risk for lyme disease by landscape connectivity, New York, New York, USA. Emerging Infectious Diseases, 25(6), 1136–1143. https://doi.org/10.3201/eid2506.181741