Mining Climate Data

Climate Change and Infectious Diseases

Kamen Shah
CU Boulder
kash1057@colorado.edu

Steve Su
CU Boulder
susc@colorado.edu

Jimmy Young
CU Boulder
jiyo4476@colorado.edu

Jacob Reed
CU Boulder
jare4857@colorado.edu

Problem Statement

There is a wealth of publications which link climate change to the increase of infectious diseases. Many studies claim that increased frequency of floods and drought due to climate change will increase the population of disease vectors such as mosquitos (LuLiang 2017). Many diseases such as Dengue, Zika, West Nile, Malaria and Yellow Fever are carried by mosquitoes. In addition, warmer climates are expanding the habitat of these vectors. For example, an increasing number of ticks have been found in higher latitudes which in the past was not a favorable environment for survival due to the cold winters (Longstreth 1989). Ticks are responsible for the transmission of Lyme disease. Every year there are more and more disease outbreaks. The World Health Organization has reported that the worldwide incidence of dengue has risen 30 fold in the past 30 years and more countries are reporting their first outbreaks. Disease outbreaks will happen regardless of climate change but climate change may be favoring the spread and survival of the disease vectors. Our study will analyze climatic and non-climatic attributes of specific disease outbreaks such as Malaria, Dengue, and Lyme disease to see if there is a common pattern. Since Lyme disease is spread by a tick and not a mosquito, we would expect the pattern of attributes to be different. By knowing the patterns which are optimal for the spread and survival of these disease vectors one may be able to predict future outbreaks in hopes of preventing the spread of disease.

Previous Findings

There are many studies which try to formulate a model to predict the influence of climate change to infectious diseases but it is rare to find a study using data mining techniques looking a wide breadth of data. Current studies are more narrowly focused with emphasis on biology of the vector and very specific data sets (Xiao 2016). However, there was one study found which does use a data mining like approach but still not in a traditional sense (Institute of Medicine 2008). Here they tracked climate attributes with respect time to predict mosquito populations.

Proposed Work

Our proposed work will be to build models for several disease outbreaks within the past 10-20 years and find common attributes. This will mainly be based on time series data and a baseline will have to be established for years when an outbreak did not occur. Since daily frequency of attributes will be too fine, we will have to bin our data in larger groups. Many studies have only included very specific data sets, in contrast we will also introduce non-climate data sets into the mix. We want to include human factors which may be a significant cause as well. From our analysis we will build a decision tree model and test it against new outbreak data sets.

Data Sets

[1] Historical Temperature Data

[a]https://www.kaggle.com/noaa/noaa-global-historical-climatology-network-daily

[b] https://www.kaggle.com/noaa/ghcn-d

These first data sets will provide us with extensive data on historical global temperatures. Using this data in conjunction with other data sets will enable us to create a very rigid model for global warming

[2] Historical Carbon dioxide levels

[a] https://www.kaggle.com/ucsandiego/carbon-dioxide

[b] https://www.kaggle.com/sogun3/uspollution

These data sets contain historical data on global pollution and carbon dioxide levels. We will use this data to complete our global warming model and furthermore determine if there are other pollutants that can correlate to the rise in infectious disease.

[3] Human health statistics

[a]https://www.kaggle.com/theworldbank/health-nutrition-and-population-statistics

[b] https://www.kaggle.com/cdc/chronic-disease

These data set will provide us with the information to see how infectious diseases have changed historically within humans. Analyzing this data will help us determine if there is any statistical correlation between these two models.

Evaluation Methods

Since the patterns we found are due to correlations we can test using cross-validation. Using a set of data that we set aside earlier which our model didn’t encounter. Unfortunately based on previous studies done at other universities just because our model may predict correctly doesn’t mean that all the features of our model found are all correlated to the prediction.

Tools

For our project we plan to leverage Python and its various libraries to conduct the data mining process. We will use the Pandas library to load the datasets and preprocess it through standardizing and cleansing the data. We will then leverage tools such as NumPy and Scipy to analyze the datasets. Finally we plan to use matplotlib to create diagrams and plots so we can easily convey our findings. Throughout our project we will collaborate using git and github so we can seamlessly share our code.

Milestones

3/15

- Setup local environment and import all required libraries
- Load data into Pandas data structure

3/22

- Merge datasets and standardize attributes
- Finish cleaning and preprocessing datasets

4/12

- Use Numpy and Scipy to analyze our datasets
- Use statistical analysis to determine if the data supports our hypothesis

4/19

- Determine if there are any patterns and whether we can extrapolate to future outbreaks
- Formulate conclusions into our final report

4/26

- Generate any diagrams that help convey our findings.
- Write up our complete findings and submit our final report

REFERENCES

[1] Institute of Medicine. 2008. Global Climate Change and Extreme Weather Events: Understanding the Contributions to Infectious Disease Emergence: Workshop Summary. Washington, DC: The National Academies Press.
P.200 https://doi.org/10.17226/12435.

[2] LuLiang, Peng Gong ,Climate change and human infectious diseases: A synthesis of research findings from global and spatio-temporal perspectives. Citation DataEnvironment international, ISSN: 1873-6750, Vol: 103, june 2017
https://www.sciencedirect.com/science/article/pii/S0160412015300489

[3] Longstreth, J. D., and J. Wiseman. 1989. The potential impact of climate change on patterns of infectious disease in the United States. In The potential effects of global climate change on the United States: Appendix G Health, ed. J. B. Smith and D. A. Tirpak for the Office of Policy, Planning, and Evaluation, U.S. Environmental Protection Agency. Washington, D.C.: U.S. Environmental Protection Agency
http://www.ciesin.columbia.edu/docs/001-488/001-488.html

[4] XiaoxuWu, YongmeiLu, SenZhou, LifanChen, BingXu, Impact of climate change on human infectious diseases: Empirical evidence and human adaptation
Environment international, ISSN: 1873-6750, Vol: 86, Page: 14-23 , January 2016
https://www.sciencedirect.com/science/article/pii/S0160412015300489

[5] https://www.who.int/globalchange/climate/en/chapter6.pdf