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 influencing 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

Scientists have concluded that "climate conditions strongly affect air-borne, water-borne, and vector-borne diseases" (HealthCom). Such findings are based on biological and environmental factors such as how climate can lengthen seasons which diseases are transmitted or how climate change can allow diseases to be introduced to un-infected regions. An example of this which does use a data mining like approach but not in a traditional sense (Institute of Medicine 2008). Here they tracked climate attributes with respect time to predict mosquito populations. In last century countless studies have surfaced which attempt predict the influence of climate change on infectious diseases, though it is rarer 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). Though there aren't many completed studies there are a couple ongoing studies. At Johns Hopkins University Applied Physics Laboratory, one proposed an approach on how

data mining could be used to predict disease frequency weeks before it happens. This would be beneficial as first responders may have a better idea of many people may be affected. Another proposed a system that would analyze a real-time stream of data which analyzed helps locate more vulnerable areas allowing us more adaptive strategies for mitigating climate change effects. Though not rooted in environmental data, data mining has been previously applied in many fields in healthcare such as, hospital resource management, infection control, treatment technique analysis, and even categorizing high risk patients.

Proposed Work

Using the CRISP-DM (Cross Industry Standard Process for Data Mining) framework we divided our task into 6 phases. The first stage included research in the area to develop a well of background information, while the second stage consisted of collecting and analyzing the data. In the third and fourth phase we will run pre-processing on the data and begin creating different models. In the final two stages we will evaluate the models and report what our models predict with a high confidence. "Applying a single data mining technique to give consistent results for all types of healthcare data may lead to predictions that are error-prone and/or very subjective to the situations" (HealthCom). Therefore, we will build multiple models for several specific disease outbreaks within the past decade or two and find common attributes shared across models. 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. The models that we create will be time series forecasting models, some with time delay, which will help root our model in time but also considering that climate may have a delayed impact on the spread of diseases.

Models which do include time delay include random forest, gradient boosting regressor, and time delay neural networks.

Motivation

Data Mining offers novel information regarding healthcare which in turn helpful for making administrative as well as medical decision such as estimation of medical staff, decision regarding health insurance policy, selection of treatments, disease prediction etc. (Divya Tomar and Sonali Agarwal).
Specifically, this study can lead to many new revelations on the catalyst to the rise of infectious diseases. Data shows that every year there is an overall increase in infectious diseases around the world, which many believe is a product of climate change. We would like to verify this hypothesis by using a large breadth of data, enabling us to determine if there is correlation with a high confidence. If we confirm this hypothesis then we can build models that can predict which areas around the world will likely be affected the worst.

Hopefully our findings can enlighten people and end the capitalistic view that America has on climate change. We also hope that our models will be successful in determining catalysts for the spread of diseases allowing people to prepare for outbreaks beforehand. For example, if our model would predict that the humidity for the next couple weeks would be extremely high in an area allowing a carrier of diseases, mosquitoes, to more easily survive; if one knew that this was going to happen beforehand one could lay poisonous bug traps in the area helping control the population.

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 it's 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. We will also utilize BiqQuery API to access some of our data. Throughout our project we will collaborate using git and github so we can seamlessly share our code.

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 climate change and its evolution in the past years.

[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 climate change models and furthermore determine if there are other pollutants that can correlate to the spread of 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. This data is crucial in helping us determine any statistical correlation between climate change and disease spread.

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

Milestones Completed

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

Milestones Todo

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

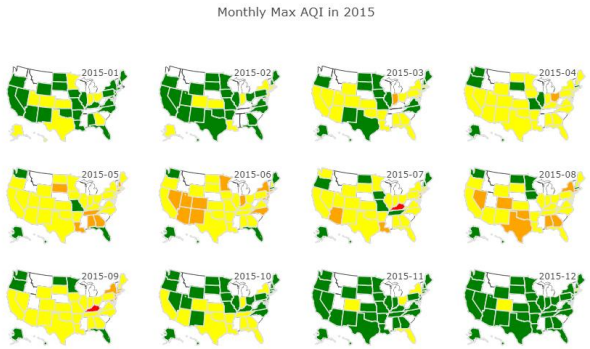
4/26

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

Results

We have already gone through the first two stages of CRISP-DM and are still working the third stage, pre-processing data, and getting our base model set up and running. At this point in our work, it is difficult to see any results or correlations as the bulk of the analysis is yet to be done. However, we can start to look at our data sets in general and see initial patterns in them. In the carbon dioxide levels set, we can see that over the last 60 years carbon dioxide levels have risen steadily. This pattern may seem basic and non interesting but it helps us to set a base as we begin to look at more complex patterns in how disease spreading and climate change relate. Looking at the pollution levels data set we can see initial patterns such as a varying level in the air quality across different months of the year. This pattern by itself is fairly arbitrary but when we begin to look at in context with disease outbreaks and how they relate to time of year and as a result pollution levels of that time of year more interesting and valuable patterns may emerge. As we progress further, it is also easy to see how we will need to account for correlations in our data sets such as carbon dioxide levels compared to air quality and temperature. Undoubtedly there is still much to be done but the preliminary work and findings are already very promising.

Graphs and Other Visualizations



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] J. A. Patz, A. K. Githeko, J. P. McCarty, S. Hussein, U. Confalonieri, Climate change and infectious diseases
https://www.who.int/globalchange/climate/en/chapter6.pdf

[6] U. Vora, A. Vakharwala, P. Chomal and M/ Sutar, "*Mining enviernmental data for prediction of transmission patterns of communicatable diseases*,"2015 17th International Conference on E-health Networking, Application & Services (HealthCom). Boston, MA, 2015, pp, 582-585.doi:10.1109/HealthCom.2015.7454569
http://ieeexplore.ieee.org.colorado.idm.oclc.org/stamp/stamp.jsp?tp=&arnumber=7454569&isnumber=7454459

[7] D. Tomar, S. Agarwal, "A survey on Data Mining approaches for Healthcare", *International Journal of Bio-Science and Bio-Technology*, vol. 5, no. 5, pp. 241-266, 2013.
http://www.sonaliagarwal.com/25.pdf