Predicting Flood Risk and Ensuing Financial Costs

Submitted For: ORIE 4741

Submitted By: Jenny Cao, Siddharth Kantamneni, Naviya Kothari

December 10, 2019

Contents

1 Problem Description 3

1.1 Climate Change 3

1.2 Consequences of Flooding 3

1.3 Our Project 3

2 Data and Data Pre-Processing 4

2.1 Federal Insurance & Mitigation Administration 4

2.2 Rainfall and Temperature 4

2.3 Sea Level Change and Elevation 4

3 Data Analysis 5

4 Conclusions 7

4.1 Next Steps 7

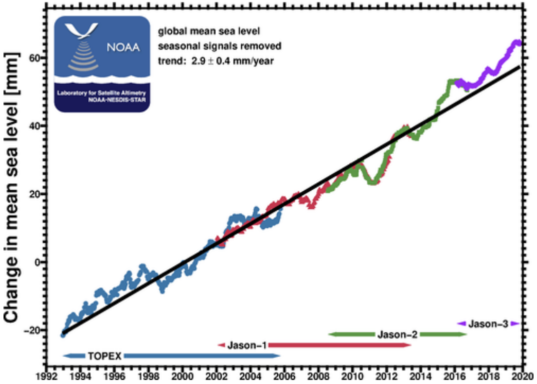
5 Appendices 8

5.1 Data Sources (Hyperlinked) 8

1 Problem Description

1.1 Climate Change

Climate change and global warming related concerns have been accelerating. Irregular weather patterns, arbitrary rainfall, untimely cyclones and snowstorms have been headlining the news notably frequently. Average precipitation in the US has increased since 1990, and it is anticipated that the northern US will experience higher winter and spring precipitation. Overall, the “trend towards increased heavy precipitation events will continue” across the entirety of the nation. Hand-in-hand with increased precipitation is the threat of sea level rise. Global sea level has risen by about 8 inches and is predicted to increase by anywhere from 1 to 4 feet by 2100. Scientists predict that in the next decade, “storm surges and high tides could combine with sea level rise and land subsidence to further increase flooding in many regions”. This threat of flooding is expected to cause cities to submerge. In fact, certain cities have already begun witnessing receding coastlines as climate change worsens.



1.2 Consequences of Flooding

Dealing with the repercussions of floods also induces a large financial cost at the state and federal levels. These costs include: 1) rehabilitating the city and 2) reimbursing victims on insurance claims. Federal Emergency Management Agency (FEMA) also uses a lot of resources after a flood in order to help the community. Planning ahead for natural disasters, specifically floods in our case, can significantly help these cities to collect resources and reorganize their finances.

1.3 Our Project

Our team is interested in exploring the ways we can predict which areas are prone to flood risk, of what magnitude, and the amount of budget they can expect to dedicate to insurance claims. As a research question, we wanted to answer: how can we predict if a city is prone to flood risk and how much budget should they assign for insurance claims?

For the purpose of this paper, we only focused on the cities in the United States to be able to completely delve into a region. We narrowed down on temperature, rainfall, elevation, and sea level rise as key predictors of floods, and therefore, built our model around these three features.

2 Data and Data Pre-Processing

2.1 Federal Insurance & Mitigation Administration

Our exploration was kindled by a Federal Insurance & Mitigation Administration's (FIMA) dataset that comprised copious amounts of time series data of past flood occurrences in the US, along with the amount of insurance claims for each. This dataset also consisted the latitude and longitude of the flood event. Given the quantum of the data, as well as its importance in our model, we decided to use the FIMA lat-long data as our common factor to accumulate rainfall, sea level change, elevation, and temperature data on.

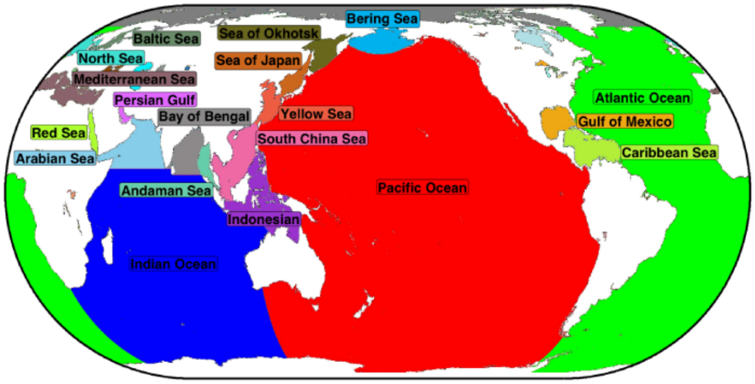
2.2 Rainfall and Temperature

It was challenging finding monthly summary statistics on any of those four variables. Fortunately, we found websites with links to daily rainfall and temperature data, dating back to the 1890s, that was collected by federally instituted stations. Each link directed us to a csv file that comprised data collected at that station only. So, for our first step, we built a web scraper that automatically downloaded around 350-375 files across the US. We also wrote a Python script that summarized rainfall and temperature for each month across 40 years at each station and combined all the stations data into one file. However, soon we realized that the lat-long pairs included in the rainfall and temperature data sets were different from the FIMA lat-longs we were interested in. In fact, we could not find climate data at all for several FIMA coordinates, which could reduce the data we had for analysis. To solve this discrepancy, we created another python script that for each FIMA lat-long, found the two nearest coordinates (or stations) available in the rainfall and temperature data respectively. Then, we averaged the rainfall and temperature values at those two stations and assigned that value to the FIMA coordinate.

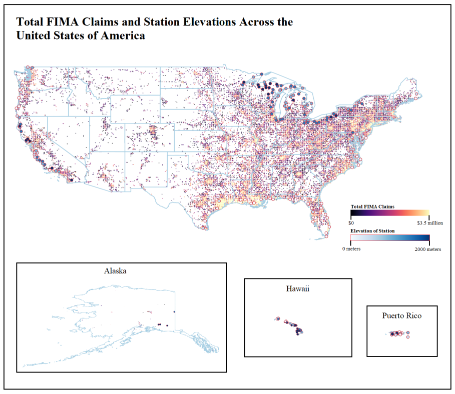
2.3 Sea Level Change and Elevation

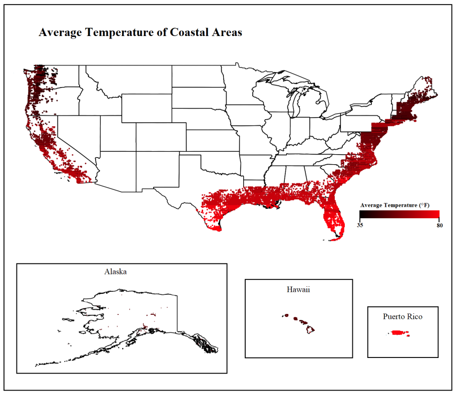
Sea level rise was difficult to obtain too. Interestingly, we found ocean level change (Fig. 1) that we re-purposed into sea level change. We did this by roughly separating US cities according to the ocean they bordered, and assuming that the ocean rise can represent sea level rise. For elevation data, we stumbled upon a source (gpsviewer) that allowed us to extract data by specifying coordinates, so we were able to incorporate this metric into our model easily.

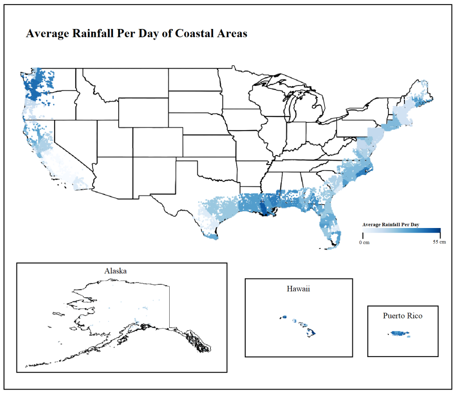
Finally, we used Excel functionalities (vlookup, predominantly) to combine all these data points across FIMA lat-longs.

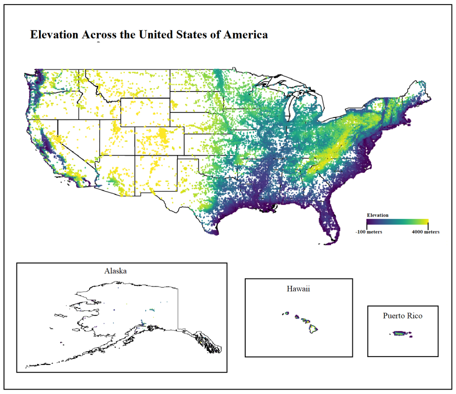


3 Data Analysis


Total FIMA Claims and Station Elevations Across the United States of America


Average Temperature of Coastal Areas


Average Rainfall Per Day of Coastal Areas


Elevation Across the United States of America

As the first step of our data analysis, we decided to use QGIS to visualize the different variables we were considering against the US map. This was with the intention of being able to visually confirm more flood prone areas and their potential correlation to higher rainfall, temperature, elevation, and/or sea level.

Going clockwise from top left, the first graph highlights how most of the highest FIMA claims happened around coastlines (eg. Texas), as one would speculate. These regions are increasingly prone to floods – be it through rainfall or cyclones. Along the same areas, elevation is also glaringly low. Simply through this visualization, one can assume that low elevation can be correlated to increased flood risks.

The second visualization shows us that areas with high FIMA claims also typically fall within low temperature zones. However, this distinction is slightly trickier to make for there is a certain fraction that falls with low temperature zones too.

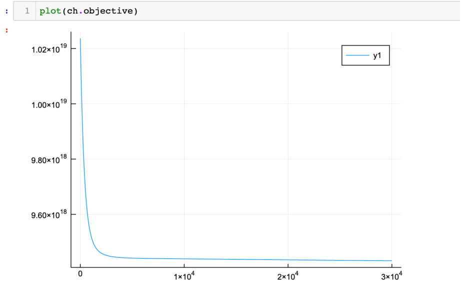
The third diagram reiterates the low elevation areas along the coast, which aligns with areas of high FIMA claims. Also, as evident, areas with high elevation also do not have any insurance claims visualizations.

Finally, the fourth diagram visualizes rainfall along the coastline of the US. Notably, there is a seemingly low correlation between insurance claims and rainfall. There are areas that experience low rainfall but still had high insurance claims and vice versa.

To analyze our data, we first demarcated the it using an 80/20 split for our project, to ensure that our model would be accurate. We also realized that we we need to change some of our data to numerical values so that the model would run smoothly. The first two feature engineering tasks we implemented were encoding the large body of water that each coastal station is closest to and adding an offset. Adding the body of water encoding makes it so that we can take into account whether or not some large bodies of water are more likely to cause flooding, or worse flooding, than other bodies. Moreover, the bodies of water were in text format, so we used real encoding to turn the values into useful numerical ones. The offset is there so that if the offset is weighted more than the other feature we have, we know that our features are not actually correlated to the output. It also increases the dimensionality of our model which in theory allows it to capture more complexities. Then, we deleted a vestigial indexing column and a column of tuples describing the latitude and longitude of the claims. These were just in the data so we could easily keep track of all the information as we were merging it all together, and they were not relevant to our output.

The first method we tried was QR factorization. We decided to use this first because, if it was effective, we would be able to quickly arrive at the correct weightings. However, this method was ineffective as it resulted in the offset being weighted more than the rest of our features. Moreover, it grossly overfit the data with the testing MSE being 10^{13} times greater than the training MSE, itself being 10^{15}.

So, we moved on to trying the Proximal Gradient Method with various regularizers and loss methods. The first, was Quadratic Loss with no regularizer whose objective value came out to 10^{18}. We then tried Huber loss, to find a balance between MAE and MSE, which output 7.3E10 as its objective value.



```
plot(ch.objective)
```

Moreover, the MSE of both the training and testing sets were similar to each other, which was encouraging. We thus concluded that Huber Loss would be more appropriate for out analysis, albeit not good per se. We then tried varying the step size and regulizers, even trying multiple level of K-Sparsity, to no avail as the objective value obstinately remained at 7.3E10, with marginal improvements in the MSE values.

4 Conclusions

Our model was able to confirm the effects of temperature and rainfall in affecting sea level change. However, we were unable to to show correlations in our features with FIMA insurance claims. This could be due to the effects of various factors that influence insurance decisions such as the robustness of the insurance system, geography, financial well-being of the state, legal support for the affected individual, and more.

Our team decided to pursue this project for its potential to have a positive social impact. We are in unison on how this model cannot be used as a Weapon of Math Destruction for it lacks any detrimental components. In fact, if successful, this model can be used to support cities in their financial planning and support individuals with their disaster preparation. As for fairness, our model could have incorporated more accurate data ranging the entirety of the US, as opposed to simply the coastlines. It is also crucial to healthily question the insurance claims data because certain groups (as distinguished by where they live) may not receive their insurance claims owing to politics. Predicting costs of floods in that case, can be prone to error.

4.1 Next Steps

Despite our final result, we believe that this continues to be a worthwhile idea to pursue. Global climate change will keep increasing most of the sea levels surrounding the US coastline. This sea level change is still going to increase flood risk for the entire country, and that will continue to cost FEMA dearly. Millions of people have been caught off-guard, lost their homes, or lost family members because we haven't been able to predict the true extent of the possibility flooding in an area. Applying machine learning to flood prediction could save lives and help FEMA prepare for disasters so that the right people get the right help.

If we were to continue with this project and work on trying improve our model, the main data set we would like to add is information on the mean price of a home in the area. It makes sense that the amount of money spent repairing the infrastructure of an area is directly related to the cost of the infrastructure itself.

Another change we would pursue would be to include all of the FIMA data we have across the entire United States of America. For this project, we only used the 8557 data points that we considered ”coastal,” however it might be interesting to compare patterns of internal and external states. With that model, distance from the coast could be included as another feature. We would also get more variation in elevation and rainfall data, which could allow us to measure the effect of a range of values on flood risks. Additionally, we would like to include general flood risk indices and those of other natural disasters. Unfortunately, most of that data is not easily available, as it is either owned by private companies and hidden away on the FIMA website, despite it being of public importance. In theory, there are API's that we could temporarily purchase access to in order to solidify our data.

5 Appendices

5.1 Data Sources (Hyperlinked)

Regional sea level time series data
Global Flood Data
Global Sea Level Change Data
FIMA Claims
Elevation Data
Temperature Data