Capstone 2 Final Report

Economic Impact of Flood Occurrence

By Kealie Pretzlav

Problem Statement

Floods occur every year all over the United States, impacting human life, local economics, and infrastructure. Climate change projections indicate that flood will likely increase in frequency and magnitude in the coming years. Flood infrastructure updates can typically only occur in areas where resources are available. Areas with higher poverty rates are typically overlooked or under-resourced, disproportionately impacting underserved populations further perpetuating poverty cycles. In this analysis, I combined a national flood dataset with US Census data on household income and public assistance amounts by zip code tabulation areas (ZTCAs) to answer the following questions:

1. Is flood occurrence a good predictor of poverty or household income?
2. How sensitive is poverty to changes in flood occurrence? Will increases in flood occurrence impact poverty?
3. Is change in flood frequency over time a good predictor of household income? I.e. is a reduction in flood occurrence an indication of community wealth for flood management infrastructure?
4. Are these datasets appropriate to answer these questions?

The goal of this analysis is to use a sensitivity analysis to quantify the impacts of changing climate on poverty status, predicting which communities will be most vulnerable to climate change-induced flood impacts, so that flood infrastructure resources can be more efficiently allocated to help the people it will benefit the most.

Datasets and Methods

To complete this project, I used three primary data sources, data from the US Census and storm event data from the National Oceanic and Atmospheric Administration (NOAA).

US Census Data

The poverty data are from the American Community Survey 5-year dataset ending in 2022 which represents 5-year average estimates for all variables from 2018 to 2022. These data are downloaded for the Zip Code Tabulation Area (ZCTA) geography. Census metrics used were household income, total population, and state fips code.

US Census ZCTA shapefiles were used to other data to the appropriate ZCTA using geopandas.

US Census American Community Survey 5-year dataset ending 2022:
https://www.census.gov/data/developers/data-sets/acs-5year.html

US Census Zip Code Tabulation Area shapefiles
https://www.census.gov/programs-surveys/geography/guidance/geo-areas/zctas.html

NOAA StormEvent

NOAA maintains a national storm event database for storm events such as tornados, wildfires, floods, hail, thunderstorms, etc. For this project, I used the “Flood” and “Flash Flood” event types to relate poverty status to flood occurrence by zip code. The data includes location (latitude and longitude, which will be mapped to ZCTAs), duration, state, date, injuries, deaths, and damage to property and crops. Additional features may be calculated such as event length (days) or other potentially relevant features as needed.

NOAA StormEvents Dataset:
https://www.ncdc.noaa.gov/stormevents/ftp.jsp

USGS National Hydrography Dataset

The USGS maintains a georeferenced dataset of streamlines for the United States at a variety of resolutions and stream orders. For this analysis I used the “medium resolution” streamlines accessed using pynhd to sum the total length of channels within a ZCTA (Figure 1). This served as a proxy for proximity or likelihood of flooding.

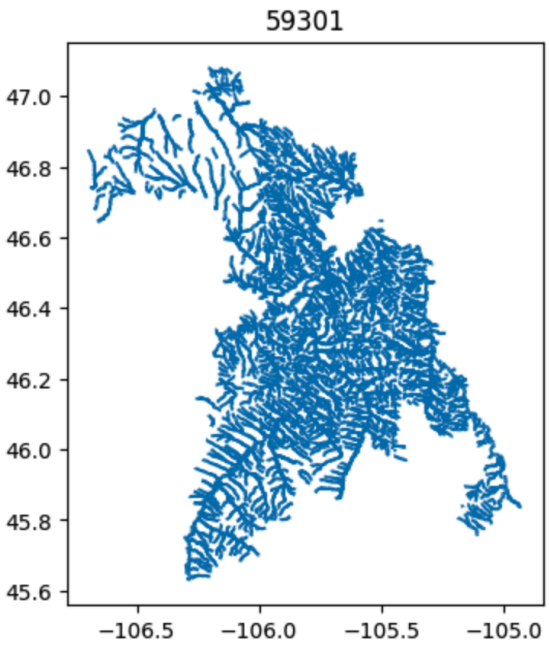

Figure 1. Example streamlines for ZCTA 59301. Axes are latitude and longitude.

Exploratory Data Analysis

The following exploratory data analysis steps were completed.

Missing data

Flood location was not consistently recorded until 2006 so had to drop floods prior to this date. Some flood locations were in rural areas outside of ZCTAs and had no associated Census data and were dropped. For simplicity, we completed this analysis for only the lower 48 states, excluding Hawaii, Alaska, and US Territories. Census data missing value was -66666666 and were removed from the analysis.

Data not used

The values of flood data features suggested that the data was incomplete or unknown. This included mostly values of zero for deaths and injuries. Many floods which occurred on the first day of the month or at midnight indicate the precise date or time may not be known. These parameters were not used.

Data Summary

Most ZCTAs have zero or very few recorded floods after 2006 (Figure 2), so data are heavily skewed. The features are very weakly correlated with each other because the full range of each feature can be mapped floods frequency less than 2 or 3.

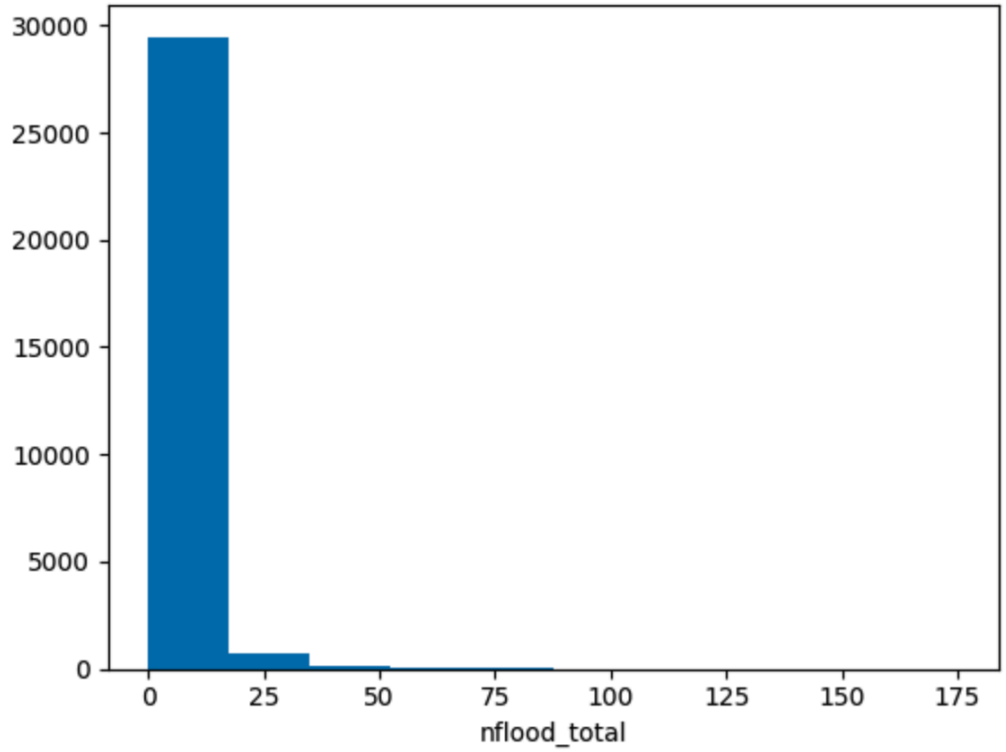

Figure 2. Histogram of total number of floods. Most ZCTAs have few floods.

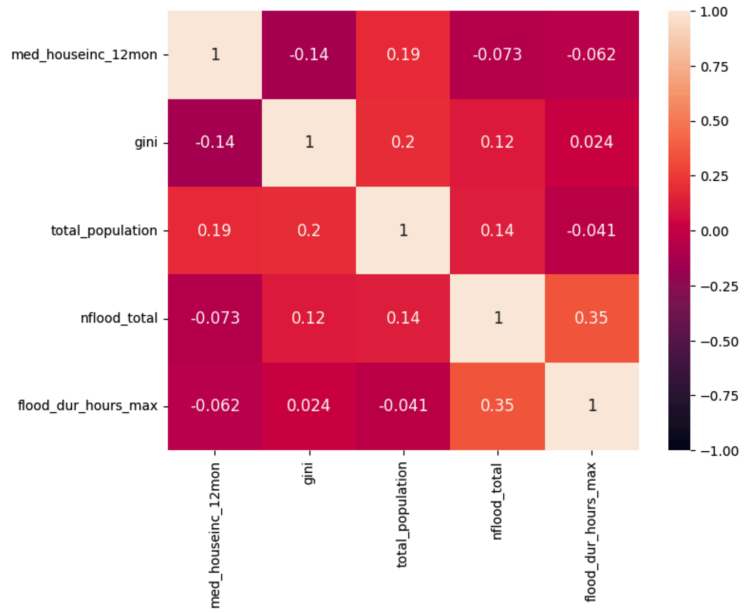

Figure 3. Correlation heatmap between features.

Feature Engineering

I first mapped the flood data to Census ZCTAs. For each ZCTA, is calculated the number of total floods that occurred since 2006, the number which in fall, winter, or spring months. Flood duration was aggregated to ZCTA by calculating the minimum, median, and maximum flood duration amongst the flood which occurred in the that ZCTA. Total channel length was calculated for each ZCTA by trimmed the NHD streamline dataset to each ZCTA shapefile, and adding the cumulative steamline distance for each polyline. I also applied a power law transform to the heavily skewed features when applicable and would help the model performance.

Modeling Methods and Results

I completed four different models. The first model used was random forest regression (RFR), which is a great first machine learning model to apply because it is fast to run, can use categorical data, and perhaps most importantly, returns feature importance in predicting the target variable. Feature importance can be used to “gut check” model results, and highlights underlying connections between features which can be connected to scientific processes when coupled with domain-specific knowledge.

The modeled target feature was median household income. The training dataset used 80% of the data. I reserved the remaining 20% as a hold-out test dataset.

The first RFR model used all input features and was re-run to remove one feature at a time with the smallest feature importance, until the R^2 value begins to significantly decrease (Figure 4). The four most important features were total population, state fips code, total number of floods, and channel length in a ZCTA. R^2 was 0.168; so the model predicted very little of the variability in household income.

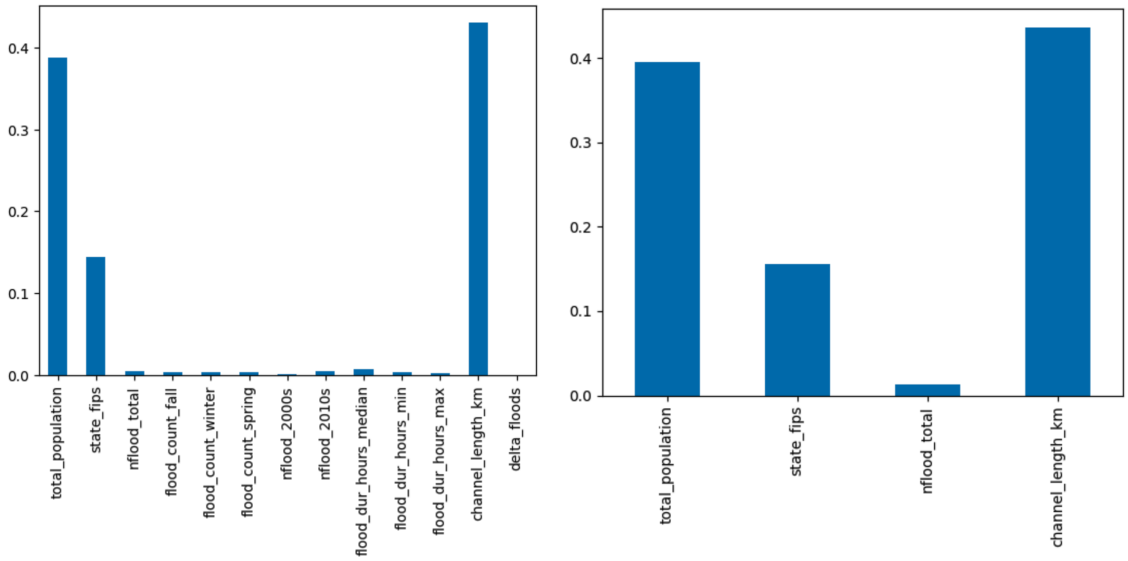

Figure 4. Feature importance of predicting the target feature, household income, for all input feature (left) and the four most important features (right).

The second model I ran was a RFR model where I transformed the features using a power law scaler function. Unsurprisingly, the results were the same as the first model as RFR is insensitive to data transformations.

For the third RFR model, I compared the previous predicted model results to the observed data (Figure 5) and saw that the model has the highest error for the lowest and highest income. This is likely because the full range of number of floods is only represented in that middle household income band, roughly from 40k to 110k per year. To complete the third model, I tried limiting the data to the ranges which had a better representation of the process I am trying to capture (economic impact of floods). So, I limited the model to ZCTAs that had more than 1 (or 10) floods, I limited the median household income to be less than 250k per year, noting that 250k seems to be the upper limit on the census questionnaire (i.e. do you make more than 250k?), and limited the analysis to ZCTA that had relatively short channel lengths within the ZCTA to avoid convoluting no opportunity to flood with economics and flood infrastructure investment.

Unsurprisingly, this resulted in much less data for the model to use and resulted in a better train $R^2 = 0.52$ but the test $R^2 = 0.14$ indicates that the model is overfitting the data and not generalized to the test dataset.

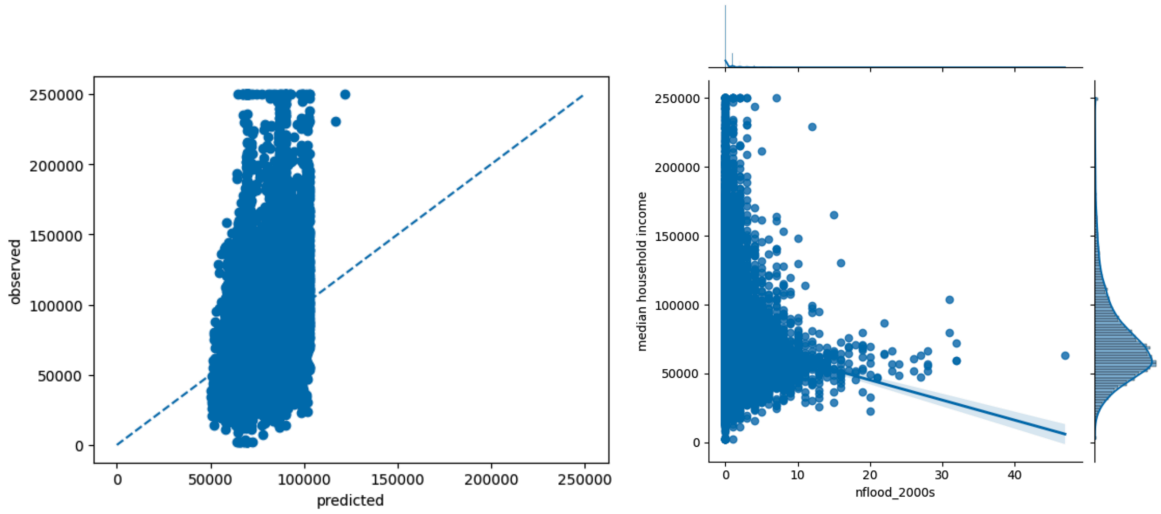

Figure 5. Predicted versus observed median household income for the second model (left) and jointplot of total number of floods versus median household income (right).

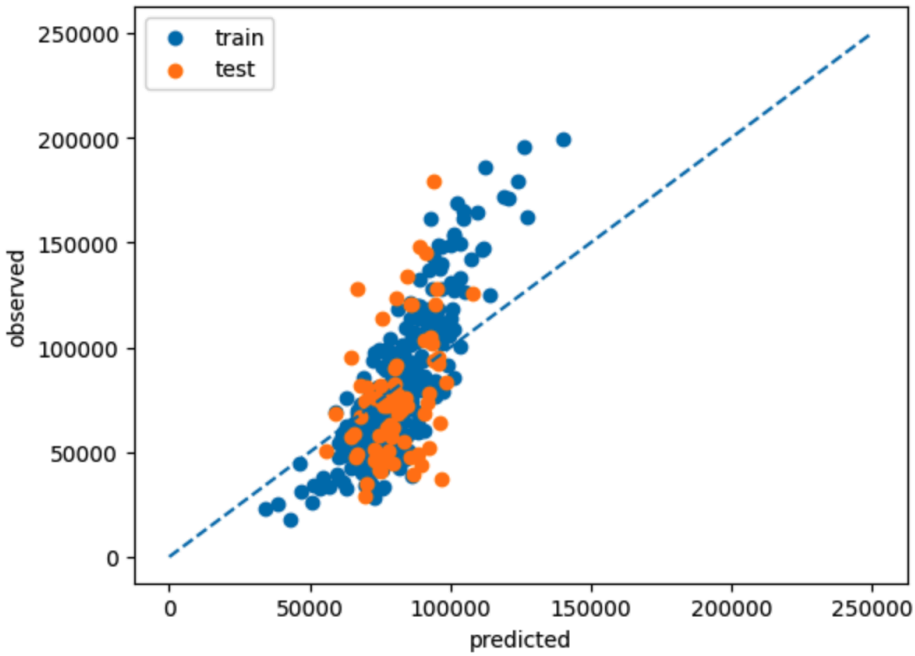

Figure 6. Predicted vs observed median household income for the filtered dataset used in model 3.

For the fourth model, I re-ran the model without the flood parameters (total number of floods). The R^2 decreased from 0.19 to 0.16 which suggests that the total number of floods in a ZCTA only accounts for 3% of the household income variability. Considering that channel length, state, and population are the other model features, 3% of model importance due to number of floods

since 2006 is a good result, especially given the generally poor model fit and relatively simple model.

I also tried a Gradient Boost model, which did improve model performance on the training dataset but overpredicted the model, revealed but considerably lower R^2 on the test dataset. I got similar results with a Neural Network model.

Next Steps

A possibly more straightforward way to answer this question is to compare flood infrastructure expenditure with household income, however I was not able to find that dataset. If it becomes available, it would be an interesting dataset to add to this analysis.

It is possible that ZCTA may be too coarse an area to pick up impacts as flooding can be very site-specific. Reducing the spatial resolution to census tract might hone in on local flood impacts, but will also spread the number of floods across the geographic units so that there is smaller numeric variability.

Finally, it is obvious that household income is a complicated metric to model and this analysis looks at only a few data sources to predict income.