Final Project Proposal

Kristina Costa

11/13/2019

The last five years have seen at least 10 natural disasters per year causing $1 billion or more in damages, and the federal government has accordingly allocated billions of dollars to help households and communities recover.Comparatively little federal money is available for communities to build resilience before suffering losses, however: FEMA's Pre-Disaster Mitigation Program had just $235 million in funding available in FY 2018. With so little funding to go around, policymakers may seek assurances that the grant program is prioritizing those communities most in need of assistance.

I propose analyzing FEMA's Pre-Disaster Mitigation Program spending along several criteria, including median area income, racial demographics, and whether grantee counties have had a federally declared disaster in the last 10 years, to help identify the characteristics of communities receiving such awards from the agency.

Data sources

My project will require using data from several federal agencies and manipulating these sources into a single, cohesive data frame for analysis. I will use FEMA's API to access information on awards made under the Pre-Disaster Mitigation Grant program and will also rely on FEMA's API to access information on major disaster declarations. Finally, I will merge the disaster and grant award data with 5-year American Community Survey estimates from the Census Bureau for affected cities, accessed via the Census Bureau's API.

Data methods

- **Manipulation** Ultimately, I am aiming to analyze data at the city-FEMA grant award level. Combining three data sets with different units of analysis will require a great deal of manipulation using `dplyr()` and `lubridate()` functions. In particular, the data on major disaster declarations from FEMA will require a lot of tidying, because FEMA has a non-chronological internal coding scheme for different kinds of disaster declarations. Disaster declarations tend to be made at the county level, but pre-disaster mitigation grants may be awarded at the city, county, or state level. Finally, Census Bureau data is collected at the census tract level and aggregated to the city (or Census-designated place) level. I expect cross-walking the county-level disaster declarations with city-level demographic data to be one of the most significant challenges of this project.
- **Visualization** This project should generate a number of interesting an informative data visualizations using `ggplot()`. In addition to generating basic histograms for data exploration, I hope to use the maps functions in `ggplot()` to illustrate the geographies

where Pre-Disaster Mitigation grants have been awarded. I will also plot the data and visualize the outcomes of my machine learning analysis.

- **Machine Learning** Finally, I will use the machine learning tools in the `caret()` and `ranger()` packages to assess whether there is a statistical relationship between three individual predictor variables—income, racial demographics, and past disaster declarations—and receipt of a Pre-Disaster Mitigation grant. I will follow best practices for subsetting my manipulated data into training and testing data sets, impute any missing values, create necessary dummy variables, and run linear regression, K-nearest neighbors, CART, and random forest models to see which perform best for each of the three predictor variables.

What does success look like?

If this project is successful, I will be able to answer questions like: Are more Pre-Disaster Mitigation Grant dollars flowing to high-income or low-income communities? What are the racial demographics of communities receiving Pre-Disaster Mitigation Grant dollars? Does it appear the grants are targeting areas at high risk for weather and climate disasters, or are the awards more random?

However, FEMA does not publish data on unsuccessful applications for Pre-Disaster Mitigation Grant awards, so I will not be able to answer questions like: Are there statistically meaningful demographic differences between successful and unsuccessful applicants? I also won't be able to assess the overall baseline characteristics of all applicants for such funds, again because the full universe of applications is not public.

Depending on what challenges I encounter in the data manipulation stage, I may need to narrow my focus to a specific geographic region, rather than all 50 states, or change my unit of analysis to the county-FEMA grant level, to make the data more manageable. Working at the county-FEMA grant level will likely make the outputs of my models less statistically significant, because county demographics tend to look more like median national demographics than do the demographics of individual cities.