SIADS 591/592 Milestone I Project Proposal

Estimating Damage from Natural Disasters in the USA

1. Team members

- Divya Badey (badeydiv)
- Sashaank Sekar (sashaank)
- Shiv Saxena (sshiv)

2. Project summary

Summarize your proposed project in a few sentences.

- What is your proposed project and why are you proposing it?
- What are the question(s) you want to answer, or goal you want to achieve?

Project Background
The National Oceanic and Atmospheric Administration (NOAA) in the USA releases the Storm Events Database that records the occurrence of storms and other significant weather phenomena having sufficient intensity to cause damage to property and/or crops. Our project will focus on exploring the types of independent variables (like population, economic activity, and associated weather statistics) that affect the total damage caused by the natural disasters in the USA. Our primary objective is to explore three major types of storms that cause widespread economic damage in the USA - tornadoes, hurricanes (flash floods and wind damage), and wildfires.

Reason for Proposal
While it is not possible to predict where and when the next natural disaster will strike, it is possible to develop models that can predict the amount of damage from such disasters. These predictive models can be useful to plan for better emergency management as the nation's increasing population grapples with the increase in both the frequency as well as the intensity of storms caused by global warming. In addition, local governments can use the models to plan for disruptions by using these models for what-if analysis.

Project Goal
Our primary goal is to analyze the various factors that contribute to the amount of damage from storms. Along with *weather related statistics such as wind speed, precipitation amount*, etc. our proposal will aim to identify other factors that can be useful to predict the amount of storm damage. Some of the factors include *population, population density, and the economic activity in a county*. We will develop visualizations to provide supporting evidence for each of these input features.
In addition, we will use SPLOMS to *determine correlation and regression fit* of the input features to the storm events' damage data. Finally, *principal component analysis* will attempt

to *find the most influential factors in predicting the damage from storms*.

3. Datasets

3.1 Primary dataset description

Describe your primary dataset. How is the data collected and how will you access it? Please share what features in the dataset are relevant to your topic. At a minimum, include the following information:

- Short description (i.e., 1-3 sentences) of its key features
- Estimated size (in records and/or bytes)
- Location (give the URL or other access method)
- Format (CSV, JSON, etc.)
- Access method (download, web scraping, API, etc.)

National Oceanic and Atmospheric Administration (NOAA)
Storm Events Database

Description: This dataset includes details on the Location (State Name, County Name, State Federal Information Processing Standards (FIPS) ID and County FIPS ID), Event Type, Damage Property, Damage Crops, Begin Date and End Date, Episode ID and Episode Narrative, Event ID and Event Narrative. This dataset has storms from January, 1950 to April, 2021. We plan to narrow down our research to focus on major storms/disasters where the damage was more than a certain threshold (e.g. $1 Million).

Size: 1.5 GB (approximately 1.8 million records)

Location: https://www.ncdc.noaa.gov/stormevents/ftp.jsp

Format: CSV

Access Method: Scrape the different csv files using Selenium script from the website.

3.2 Secondary dataset(s) description

Describe your secondary dataset(s). How is the data collected and how will you access it? Please share what features in the dataset(s) are relevant to your topic and describe the data types you're expecting. At a minimum, for each secondary dataset include the following information:

- Short description (i.e., 1-3 sentences) of its key features
- Estimated size (in records and/or bytes)
- Location (give the URL or other access method)
- Format (CSV, JSON, etc.)
- Access method (download, web scraping, API, etc.)

In order to explore various weather and economic factors that will impact the damage from natural disasters, we will need to access several different secondary datasets. These are:

1. US Census.gov

1. https://www.census.gov/data/developers/data-sets/cbp-nonemp-zbp.html
   Description: The County Business Patterns dataset will provide the number of businesses and their total payroll in the county for the year previous to when it was affected by the storm. Similarly, the Non-Employers dataset will provide the number of small businesses and the total revenue generated by them in the year previous to the occurrence of the storm.

2. https://www.census.gov/data/developers/data-sets/popest-popproj.html
   Description: The Census Bureau's Estimate Program dataset will provide the population estimate in the county affected by the storm for the year previous to the occurrence of the storm.

3. https://www.census.gov/data/developers/data-sets/economic-census.html
   Description: The Economic Census (conducted every 5 years) dataset provides the economic activity across major industries in the county affected by the storm nearest to the year of the occurrence of the storm.

All three data sources will provide data for the input features of the models for estimating the damage caused by the three types of natural disasters researched in our study.

Size: 1 record for every county in the storm events dataset

Location:

1. https://api.census.gov/data/2017/cbp?get=NAME,NAICS2017_LABEL,ESTAB,PAYANN,EMP&for=county:201&in=state:48&key=USER_GENERATED_KEY
2. https://api.census.gov/data/2017/nonemp?get=NAME,NAICS2017_LABEL,NESTAB,NRCPTOT,&for=county:201&in=state:48&key=USER_GENERATED_KEY
3. https://api.census.gov/data/2017/ecnbasic?get=NAICS2017_LABEL,NAICS2017,ESTAB,FIRM,EMP,RCPTOT&for=county:201&in=state:48&key=USER_GENERATED_KEY

Format: CSV

Access Method: Download using the census.gov API.

2. U.S. Drought Monitor (USDM)

https://www.drought.gov/data-maps-tools/us-drought-monitor

Description: We will use the USDM's drought severity data for the county or the zone where the wildfire event occurred as found in the storm events dataset. This will be used as input to the model for predicting the damage from wildfires.

Size: 1 record for every county or zone in the storm events dataset where the event type is a wildfire.

Location: https://droughtmonitor.unl.edu/DmData/GISData.aspx

Format: CSV

Access Method: Scrape the different csv files using Selenium script from the website.

3. Weather.gov

https://www.weather.gov/gis/ZoneCounty

Description: The primary dataset is missing the latitude and longitude of storm events that occur in a zone (a National Weather Service zone covers more than one county). We use the zone county correlation dataset to get the latitude and longitude of a given zone FIPS ID for an event in the storm events dataset. In addition, the county FIPS for the counties in the zone is also provided.

Size: 1 record for every county in the storm events dataset

Location: https://www.weather.gov/source/gis/Shapefiles/County/bp10nv20.dbx

Format: CSV

Access Method: Download directly from website

4. National Climatic Data Center (NCDC)

https://www.ncdc.noaa.gov/cdo-web/

Description: We get the list of weather stations around a given county FIPS ID. We intend to use the list of weather stations to find the nearest weather station to an event from the storm events dataset.

Size: Varies depending on the result of the query for the county FIPS ID from the storm events dataset.

Location: https://www.ncdc.noaa.gov/cdo-web/api/v2/stations?locationid=FIPS:08049

Format: CSV

Access Method: Download using the NCDC API

5. National Center for Environmental Information (NCEI)

https://www.ncei.noaa.gov/support/access-data-service-api-user-documentation

Description: The daily summaries dataset will provide the daily weather statistics (minimum and maximum temperature, precipitation, snowfall) for the weather station nearest to the storm's event location. We use the precipitation, snowfall and temperature data to calculate their rolling average over the last twenty years. This information will be used as input for the model for estimating the damage caused by wildfires.

Size: Approximately 7300 records per weather station nearest to a county FIPS in the storm events dataset.

Location:
https://www.ncei.noaa.gov/access/services/data/v1.?dataset=daily-summaries&stations=USC00397277&startDate=2000-01-01&endDate=2020-12-31

Format: CSV

Access Method: Download using the NCEI API

3.3 [✔] Affirm: datasets are public.

Please check the above box to confirm that your primary and secondary datasets are accessible and available to your classmates and the instructional team.

4. Cleaning and manipulation

How will you join your primary and secondary datasets? What cleaning and manipulation challenges, if any, do you anticipate?

We plan to join the primary and secondary datasets using the **state FIPS** and the **county FIPS ID**. In order to ensure successful retrieval and plotting of information, we plan to employ a number of data cleaning and preprocessing steps. These include filtering out rows that have missing FIPS ID information, selecting records from the 50 US states and Washington DC, and filtering out rows that have damage greater than the threshold.

The storm events dataset has events listed either under a county or a National Weather Service zone where a zone covers multiple counties. We can get the county FIPS for each county in the zone from the weather.gov dataset as described in the secondary dataset section. However, **one challenge** is how to split the cost of a storm damage across multiple counties, so that we can plot the damage on the map using a choropleth map.

In the storm events dataset, whenever the event occurs in a National Weather Service zone, **the latitude and longitude is missing**. We need the geolocation so that we can identify the nearest weather station. In order to solve this problem, we have to get the latitude and longitude of the county from the weather.gov dataset as described in the secondary datasets section.

During our initial analysis, we have seen situations where the weather station nearest to a county FIPS ID has **missing weather statistics** in the reported daily summary. To solve this problem, we will have to expand our search to the next nearest weather station.

5. Analysis

Describe any analyses you plan to undertake. For each, please give the technique or approach and briefly explain what you expect to learn from it.

We plan to **split** the primary NOAA storms dataset by the type of storms (tornadoes, hurricanes, wildfires) and filter only those storms where the damage was over a certain threshold (e.g. $1 Million).

Once we have split the data by type of storms, our strategy will involve **collecting the data for the input features specific to each type of storm**.

For **tornadoes**, the main input features set about the tornado will come from the storm events dataset (e.g. tornado strength, length and width, magnitude of wind speed). In addition the census.gov dataset will provide information about the local economy in the county affected by the tornado.

For **hurricanes**, the location and duration of the event will come from the storm events dataset. We will get the weather related information (mainly precipitation to factor for ground saturation for flood damage) from the weather stations found in the NCDC dataset and the daily summaries for each weather station from the NCEI dataset.

Lastly for **wildfires**, the location and duration will come from storm events dataset and the drought information will be fetched from the USDM dataset. In addition, the weather related information (precipitation, snowfall) will be fetched from the weather stations found in the NCDC dataset and the daily summaries for each weather station from the NCEI dataset.

6. Visualizations

Describe in 1-3 sentences at least **two** data visualizations that you plan to create. Include the chart type (e.g. bar chart, scatterplot, SPLOM, etc.) as well as the variables (features) you intend to plot.

- **Bar chart**
  We will use bar charts to plot the frequency of the 3 types of storms (tornados, hurricanes, and wildfires) for every year for which we have the data. We also plan to plot the total damage (after adjusting for inflation) caused by the 3 types of storms for every year for which we have the data.

- **SPLOM**
  We plan to plot all input variables along with the storm damage to see if any obvious correlations exist between the chosen input variables for each of the 3 types of storms researched in our study. This will provide us with an opportunity to identify confounding variables and allow us to control for their impact on our analysis.

- **Geographic visualizations**
  We plan to use choropleth maps showing the total damage from storms in the 50 US states and Washington DC using color as the retinal variable. We can add animation to plot the damage for every year for which we have the data.

7. Ethical considerations

Does your choice of data raise any ethical issues? If so, briefly describe the concern and how you plan to mitigate it.

1. Our coverage is limited to specific storm types with damage greater than a specific threshold. This may lead to data science ethical issues where not all storms are represented. Our findings about storm damage relationships may only work for these specific storm types but not all.
   Mitigation: Our report will include a clear statement about the scope of our study.

2. Our analysis of storm damage includes a limited number of data sources and a limited number of input features. The scope of our project means that we will have significant gaps in our data and our final conclusions cannot be comprehensive.
   Mitigation: Our report will include a section on the assumptions made in our analysis.

3. Our data is publicly available from government APIs. We will need to be cautious when using these APIs so that we do not violate the access terms of use such as sending multiple requests within a given timeframe.
   Mitigation: We will download the data once and save it for future use.

8. Contributions

Indicate the contribution that each team member will make to the project.

We currently plan to divide the tasks evenly, but we will be flexible and make adjustments as needed depending on the availability of the team members.

Divya

- Data source research and selection, data cleaning, exploratory data analysis
- Visualization
- Report writing

Sashaank

- Data source research and selection, data cleaning, exploratory data analysis
- Visualization
- Report writing

Shiv

- Data source research and selection, data cleaning, exploratory data analysis
- Visualization
- Report writing