To: Professor X

From: Aaron Jay, Carolyn Shi

Subject: Midterm Report for Predicting Storm Casualties

Date: November 7, 2019

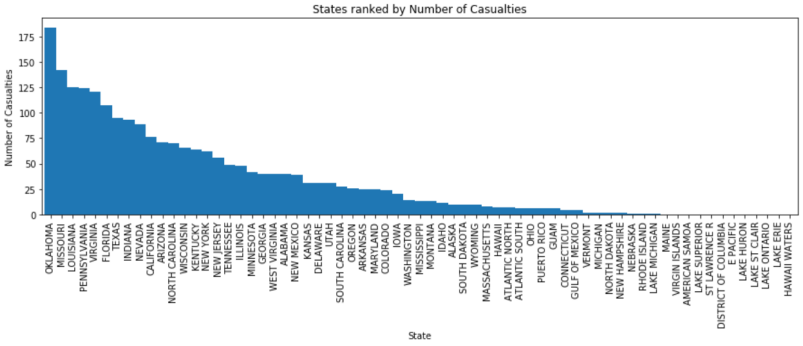
---

1 Dataset History and Description

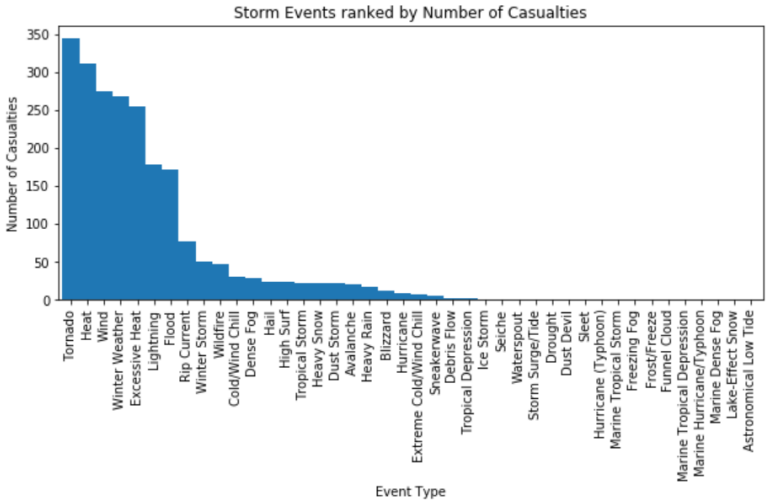
- This dataset is provided and maintained by the National Weather Service (NWS) and contains information about storms dating back to 1950. In the beginning, only data about tornadoes was collected, until 1955 when information about Thunderstorm Wind and Hail were added from paper publications. Then in 1996 the event types were expanded to what exists in the database now of 48 different storm types.
- The later addition of these event types has also affected the features that have been collected about storm events. Thus, we are choosing to focus our attention on storm events that have had more descriptive data collected.
- In the raw data, there are 50,000-60,000 examples for each year of storm event details (though there are significantly fewer examples with fatalities). The features of the data can be grouped into the following categories:

Name	Description	Features	Data Types
Date/time	Date and time when a storm event occurred.	5	Date/time
Geographic Details	Associated terrain of a storm event location based on the International Geosphere–Biosphere Programme (IGBP) land classification.	1	Nominal
Event Type	NWS classification of weather events among 48 available types.	1	Nominal
Casualties and Damages	Counts of injuries, fatalities, and economic damage estimates.	6	Real
Event Characteristics	Measurements taken from a storm event.	5	Real, Ordinal, Nominal
Narratives	A textual account of developments during a storm event. Includes measurements of weather conditions and accounts of fatalities.	2	Text
Fatality Details/Features	Details about a storm event victim, and circumstances of death.	4	Real, Nominal

- The International Geo-sphere–Biosphere Programme (IGBP) land classification was a separate data source that we included to give more context to the raw geographical data in terms of longitude and latitude given with the NWS datset.
- We investigated two factors for casualties in the dataset: location (state) and event type.



Since there is a skewed pattern towards certain states having casualties, we use this as a basis for including more geographic data in the future. For preliminary analyses, we decided to focus on storm event types and their effects on casualties.



As the graph shows, the bulk of the casualties are due to a handful of storm event types. In conjunction with our prior justification regarding the history of the data, our regression model will focus on tornadoes, floods, winds, and hail type storm events.

- The main dataset is actively managed and updated by the National Weather Service, thus it is well-formatted and well-documented. We have not had or anticipate huge issues with corrupted data.
- In the data many fields will be missing due to the fact that the column is unrelated to the event recorded (e.g. column `FLOOD_CAUSE` for event type of "Tornado"). In such cases, we will attempt to break up the dataset into rows where those columns would make sense.

---

2 Model Development

- Testing, handling over (and under-)fitting: We plan to randomly sample from the original dataset to test the models we develop. It is unlikely that our models will over fit since we are controlling the features that go into the model for the sake of interpretability. Nevertheless, we will address any overfitting concerns through the use of regularizers. Our more immediate concern is handling underfitting by incorporating features extracted from other datasets, and from un-utilized data in the original dataset such as the event narratives.

- For our initial model, we chose to use simple linear regression since predicting casualties is a regression problem. We first used continuous features in order to compute correlations, though this attempt proved to be too limited due to the small number of features used. We then included categorical variables such as event type, but there was not much improvement in performance. This is due to the fact that a majority of the storm events recorded had no deaths or injuries.

---

3 Future Developments

- Feature Engineering:
  - One Hot Encoding: We plan to use one hot encoding for categorical variables such as Flood Cause and Land Cover to transform it into a real value for use in a regression model.
  - Text Parsing: We plan to extract more features from the narratives that are provided in the data. One example of this is "episode type", where individual storm events may be a part of broader phenomena (e.g. hurricanes).
- While the NWS dataset does provide measurements for events such as tornadoes, we also hope to find data regarding other high-casualty events like lightning and heat waves.
- We will look into the possibility of adding in more features such as population density to help with under-fitting the model.
- Using different models: while predicting the number of casualties is a regression problem, our initial attempt showed that linear models were not necessarily good for our data. Thus, decision trees may also be a good consideration for sorting through zero-casualty events. Furthermore, developing more specific models based on types of storm events may provide more context for imputing missing data.