Data Analysis of Severe Weather Events

Jason Kang

Aaron Talley

Michael A. Esparza, PMP, PgMP, CPEM

Data Analytics, IST 707

Dr. Ami M. Gates

14 September 2019

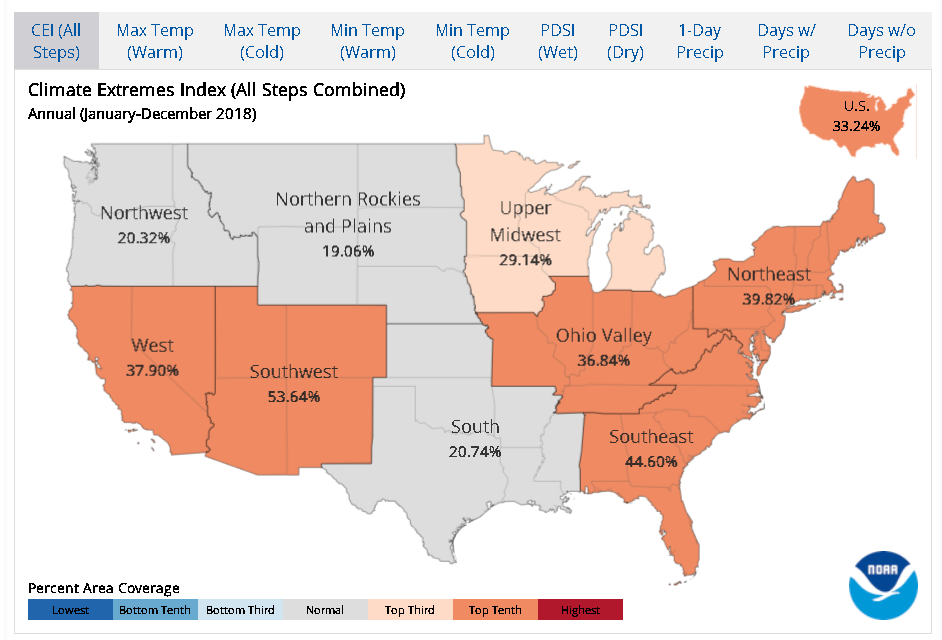
**Introduction**

Severe weather events have increased in recent decades (National Centers for Environmental Information (NCEI)). Some of them, severe weather events such as hurricanes, typhoons, drought and wildfires, are no longer uncommon (National Centers for Environmental Information (NCEI)). Each year, millions of lives are affected: and not just in the number of fatalities, but in real danger to human life and limb.

In addition to the placing human lives, severe weather is the direct cause of billions of dollars’ worth of property damage; and precious crops are ruined or rendered unusable, especially in the heart of America’s breadbasket. Viewed over time one can infer that the cycle of severe weather events gets worse with each year (see fig. 1). Severe weather is affecting how and where people live. It is a complex issue that cannot be taken lightly, and those leaders that have the duty and responsibility to protect citizens should have a knowledge of the impact of these events in order to facilitate emergency management, medical priorities, and assignment of resources and personnel when responding to severe weather events.

Presented is an analysis of severe weather events data in support of decision making for those critical stakeholders involved in preparing and planning to protect human life and mitigate the risks of severe weather to the U.S. economy . It is a review of the available data (National Weather Service) and addresses two critical questions that most influence the decision maker’s deliberate and rapid response planning process:

* Which severe weather events are most harmful to population health?
* Which severe weather events have the greatest economic consequences?



**Figure 1. Climate extremes visualized for the contiguous U.S. NOAA National Centers for Environmental**

**information, Climate at a Glance: National Time Series, published September 2019, retrieved on**

**September 13, 2019 from https://www.ncdc.noaa.gov/extremes/cei/regional-overview.**

**Analysis and Models**

**About the Data**

The data is originated from National Oceanographic and Atmospheric Administration (NOAA), the authority of national weather data (National Weather Service). It is an accumulation of weather information from year 1950 to 2011, for a total of 61 years. The data contains 902,297 data points, and each data point consists of 37 unique fields. The following is a data dictionary of the important fields (see table 1).

|  |  |
| --- | --- |
| **Severe Weather Events Data Set Dictionary** | |
| **Fields Concerning Time** |  |
| ***Observation*** | ***Data Attribute*** |
| **\*BGN\_DATE\*** | **\* start date of the event** |
| **\*BGN\_TIME:\*** | **\* start time of the event** |
| **\*END\_DATE:\*** | **\* end date of the event** |
| **\*END\_TIME:\*** | **\* end time of the event** |
| **Fields Concerning Damage Magnitude** |  |
| ***Observation*** | ***Data Attribute*** |
| **\*FATALITIES:\*** | **\* number of deaths** |
| **\*INJURY:\*** | **\* number of injuries** |
| **Field Concerning Event Size** |  |
| ***Observation*** | ***Data Attribute*** |
| **\*BGN\_RANGE:\*** | **range of the event (in tenth of a mile) at the beginning** |
| **\*END\_RANGE:\*** | **range of the event (in tenth of a mile) at the end** |
| **Fields concerning event type** |  |
| ***Observation*** | ***Data Attribute*** |
| **\*EVTYPE:\*** | **type of event (e.g. HAIL, TORNADO, FLASH FLOOD, HEAVY SNOW...etc)** |

**Table 1. Severe Weather Events data Set Observations and Attributes. Source: Severe Weather Events Data Set.**  **National Centers for Environmental information, NOAA, published September 2019, retrieved on September 13, 2019 from https://www.ncdc.noaa.gov/cdo-web/datasets.**

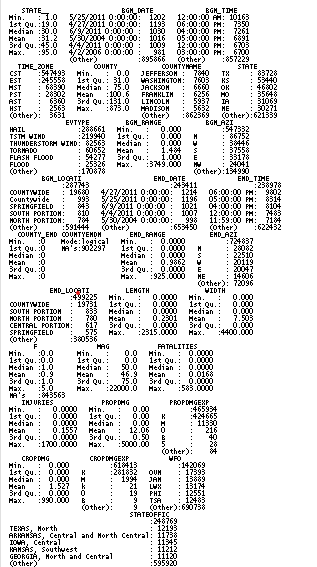
**Visual and Statistical Exploratory Data Analysis (EDA)**

Visual EDA presents the following structure *(str(model\_data)* for the Severe Weather Events Data Set (see table 2).

|  |  |
| --- | --- |
| **Severe Weather Events Data Set Structure** | |
| **Data Frame** | **902276 obs. of 8 variables:** |
| **$ EVTYPE** | **Factor w/ 985 levels " HIGH SURF ADVISORY”, 834 834 834 834 834 834 834 834 834 834 ...** |
| **$ STATE** | **Factor w/ 72 levels "AK","AL","AM”, 2 2 2 2 2 2 2 2 2 2 ...** |
| **$ STATE\_\_** | **num 1 1 1 1 1 1 1 1 1 1 ...** |
| **$ INJURIES** | **num 15 0 2 2 2 6 1 0 14 0 ...** |
| **$ FATALITIES** | **num 0 0 0 0 0 0 0 0 1 0 ...** |
| **$ PROPDMG** | **num 25000 2500 25000 2500 2500 2500 2500 2500** |
| **$ CROPDMG** | **num 0 0 0 0 0 0 0 0 0 0 ...** |
| **$ TOTAL.DMG** | **num 25000 2500 25000 2500 2500 2500** |

**Table 2. Severe Weather Events data set structure. Source: Severe Weather Events Data Set.**  **National Centers for Environmental information, NOAA, published September 2019, retrieved on September 13, 2019 from https://www.ncdc.noaa.gov/cdo-web/datasets.**

Statistical EDA outputs the following quantitative Objective Quality Evidence (OQE) (see fig. 2).



**Figure 2. Statistical EDA of the Severe Weather Events data frame.**

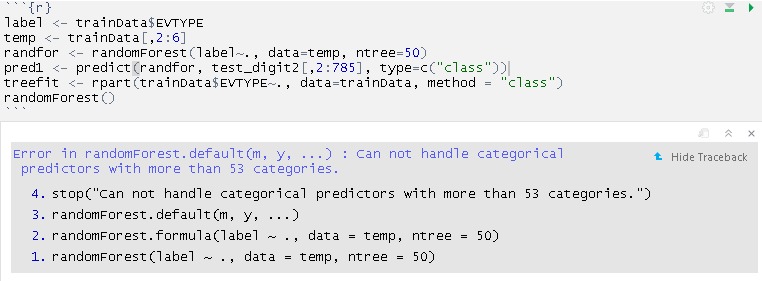
**Cleaning and Prep**

Performing “na.omit” for complete classes was required to pre-process the data set for follow on analysis of the objective research task. After initially loading data.frame, there were some observation that contained zero values. This was realized that the observation did not apply, but that there were no fatalities, loss of life and limb, or property/crop damage within the observations. While the results returned would be initially construed as an empty data frame, in cleaning and preparing the data, it was decided to execute the damage calculations and provide a more rigorous and relevant data frame to meet the research objective.

**Models**

**Decision Tree/Random Forest**

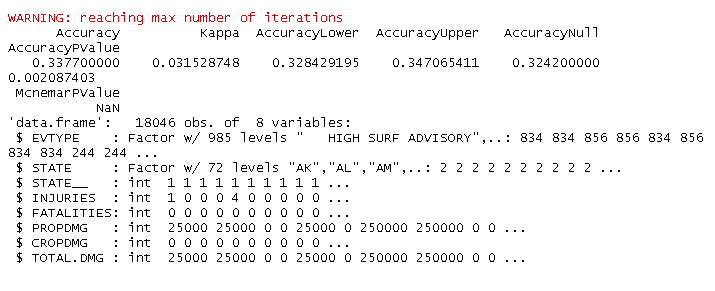
One of the weaknesses of the Random Forest and by abstraction Decision trees algorithms was discovered when attempting to apply the model to the research data set. Since there are more than 53 possible labels, Random Forest Fails to run And Decision Trees have an incredible long run time (over 6 hours). So Random Forest and Decision trees have a negative performance measure of effectiveness (MOE) when dealing with labels measuring in the hundreds. This is a Lesson Learned and further experimentation to adapt Decision Tree/Random Forest to this type of data set is necessary.



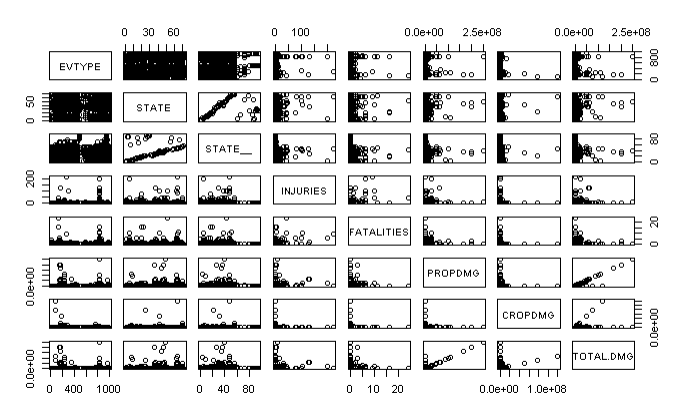
**Figure 3. Decision Tree and Random Forest results when applied to the Severe Weather Events Data Frame.**

**Support Vector Machine (SVM)**

SVMs with nonlinear kernels add additional dimensions to the data in order to create separation in this way. Essentially, the kernel trick involves a process of constructing new features that express mathematical relationships between measured characteristics. For instance, the altitude feature can be expressed mathematically as an interaction between latitude and longitude— the closer the point is to the center of each of these scales, the greater the altitude. This allows SVM to learn concepts that were not explicitly measured in the original data (Lantz). The SVM performed adequately to present the statistical features of the data set and produced better MOE than with the Decision Tree/Random Forest Model (see figs 4 and 5).



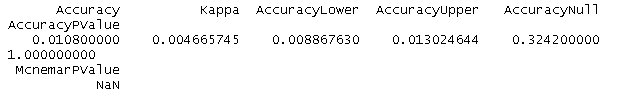
**Figure 4. SVM Statistics of the Severe Weather Events Data Set.**

****

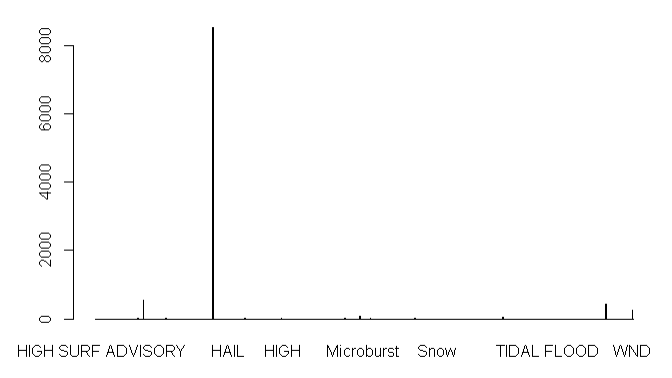
**Figure 5. SVM Visualization of the Severe Weather Events Data Set.**

**Naïve Bayes.**

Regarding this data analysis and the base data set, the Naive Bayes model performed poorly, producing results with the worst accuracy of the available models employed (see fig. 6 and 7). In reviewing the literature, the most probable cause is that The Naive Bayes classifier is often used for text classification (Lantz). This was not he case with the Severe Weather data set where text classification was necessary.



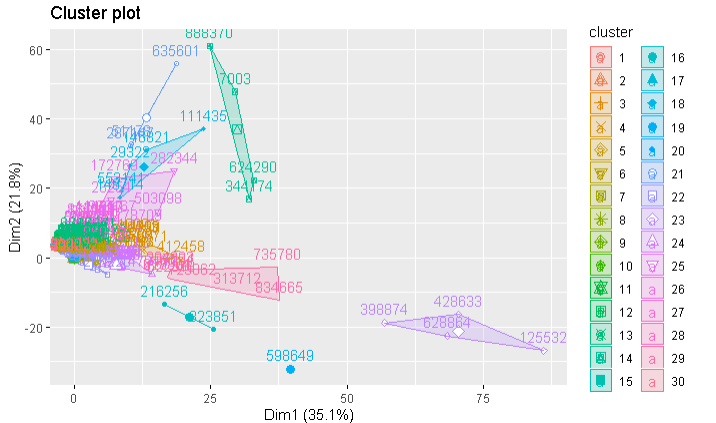
**Figure 6. Naïve Bayes Statistics of the Severe Weather Events Data Set.**



**Figure 7. SVM Visualization of the Severe Weather Events Data Set.**

**K Means**

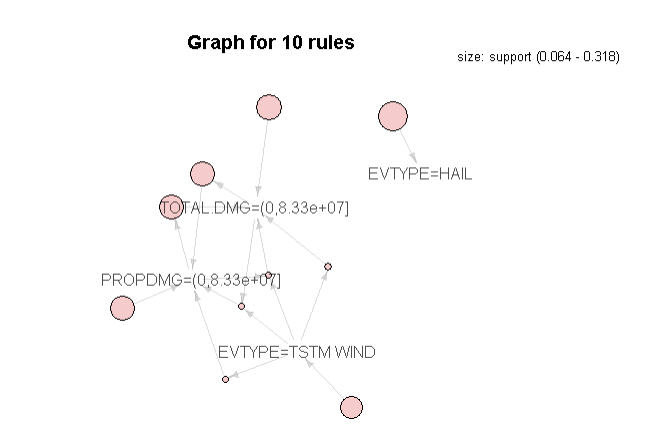
K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labelled, outcomes (Garbade). In the case of the data set under analysis, increasing the number of clusters decreases the total value of the accuracy within the data set and analysis results (see fig. 8).



**Figure 8. K Cluster Visualization of the Severe Weather Events Data Set**

**Association Rule Mining (ARM).**

The ARM model produced the best results in the analysis of the Severe Weather data set and produced the most informative results. It is this model that can best convey to a lay audience those severe weather events that produce the most negative impact and facilitates high level decision making and planning.



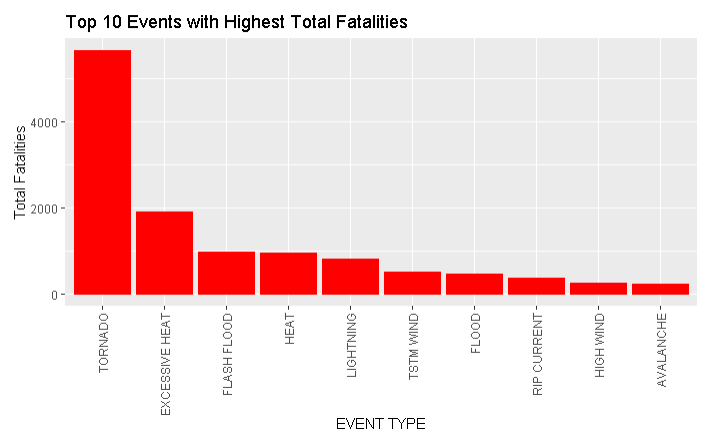
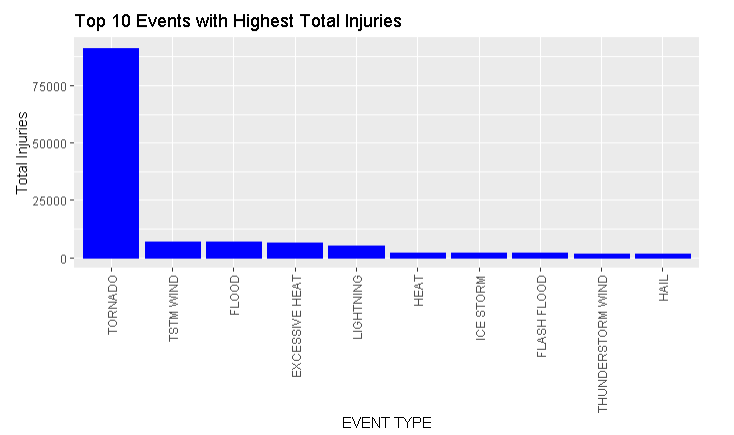
**Figure 9. ARM Visualization of the Severe Weather Events Data Set.**

**Results**

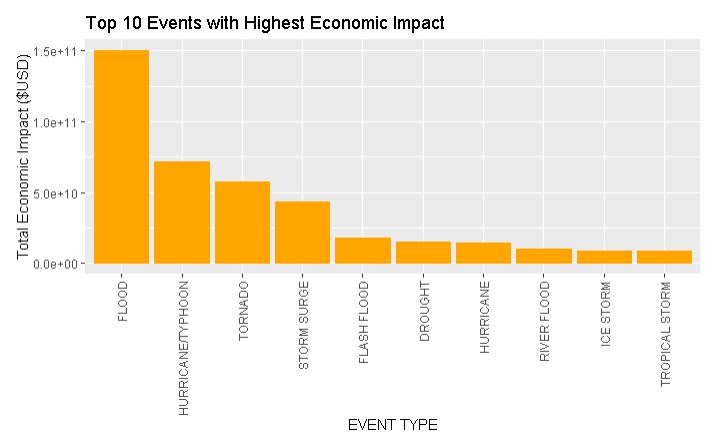
Based on the final data analysis and models, the following results (see figs, 10,11,12) are presented to address the key questions that are the basis for this research:

* Which severe weather events are most harmful to population health?
* Which severe weather events have the greatest economic consequences?

Upon review of the presented analysis, the Severe Weather Event Type (EVTYPE) Tornados is the leading cause of injuries and loss of life and limb to the U.S. population. The Severe Weather event with the greatest impact on the U.S. Economy is flooding.



**Figure 10. Highest Total Fatalities.**  **Figure 11. Highest Total Injuries.**

****

**Figure 12. Highest Economic Impact Injuries.**

**Conclusions**

Being in a position of leadership, with duties and responsibilities that encompass the safeguarding of lives, property and resources, requires quick and accurate access to data that prescribes, not merely describes, informs or predicts. As is often the case, our elected leaders are not data scientists; and most often do not carry a science degree. Their focus is politics and public policy.

Being a data scientist requires the ability (and some would say the art) of speaking truth to power; and not spinning a data analysis to any particular policy or agenda. This is often a difficult undertaking - even though we are data scientists and use our multidisciplinary approach in accordance with the scientific method, we are after all, human, and are all subject to decision bias, personal bias, and emotional reactions. The challenge as data scientists, then, is to present our results as the data analysis outputs, whether we agree with the results or not. It is not a matter of agreement, it is a matter of scientific truth based on data, information and as stated, the scientific method.

This place upon the data scientist the moral and ethical responsibility to ensure that we report our findings completely and accurately. The peer review process is also critical to our endeavors to make the world a better place via our work. With each data scientist watching the others back, we can only improve upon the research that we undertake. In the specific regime undertaken in this research assignment, we have come to that junction where public policy and science meet. It is for the greater good that we attempt to convey our findings in a manner that is easily understandable to the lay person. As our professor stated, decision makers do not want to know about which data science techniques we used to arrive at our conclusions, they just want to know what they are looking at when presented with a graph and what our recommendations are so that they can make informed decisions.

In the end, as data scientists we do influence what occurs at the public policy level. The more we gain the trust of our elected officials by virtue of our work, the more we be able to serve the public good, which is always a noble endeavor. Looking at severe weather events and informing our leaders of what the real threats are regarding severe weather events, and not anecdotal information, will save more lives and protect property and resources. We have then done our moral duty as scientists and citizens and can be rightly proud of our chosen profession.

# Works Cited

Garbade, Dr. Michael J. *Towards Data Science*. 12 September 2018. Web Page. 13 September 2019. <https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1>.

Lantz, Brett. *Machine Learning with R, 2nd Edition: Expert Techniques for predicitve modeling to solve all your data analysis problems.* Kindle. New York: Packt Publishing, 2019. Book.

National Centers for Environmental Information (NCEI). *Extreme Events*. 2019. NOAA Website Page. <https://www.ncdc.noaa.gov/climate-information/extreme-events>.

National Weather Service. *Directives*. 2019. Web Page. 13 September 2019. <https://www.nws.noaa.gov/directives/>.