Severe Weather Events Causing Economic Damage or Casualties in the United States

Cameron Charness

September 25, 2018

```
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
options(scipen=1,digits=2)
options(knitr.table.format = "latex")
library(formatR)
library(dplyr)
library(tidyr)
library(R.utils)
library(data.table)
library(lattice)
library(latticeExtra)
library(kableExtra)
library(rlang)
```

Executive Summary

For this report, we analyze data from the National Oceanic and Atmospheric Administration's storm database from 1950 to 2011. This database contains observations from various types of severe weather events across the entire United States (and some marine areas), including reports of property and crop damage as well as injuries and fatalities resulting from each event. We cleaned up this data to include only events resulting in either deaths, injuries, or significant damage. We then consolidated the large number of event categories into a few broader event types in order to analyze their impacts. Our analysis indicates that there is some difference between which types of events cause the most economic damage and which cause the most casualties (including both injuries and deaths), though there is of course substantial overlap. Ultimately, we conclude that casualties are most likely to occur in the wake of more localized, less predictable events such as tornadoes and thunderstorms, whereas economic damage comes more from larger or longer-duration events such as hurricanes and flooding.

Data Processing

We begin by unzipping and extracting the raw data into a table with read.csv, and using the head function to take a quick look at the first few rows. We'll want to use the cache option on this code chunk, as the dataset is quite large and takes nontrivial time to load into R.

```
bunzip2("repdata_data_StormData.csv.bz2", "repdata_data_StormData.csv",
    skip = TRUE)
```

```
## [1] "repdata_data_StormData.csv"
## attr(,"temporary")
## [1] FALSE
```

```
rawdata <- read.csv2("repdata_data_StormData.csv", header = TRUE,
    sep = ",")
head(rawdata)
```

```
##   STATE__           BGN_DATE BGN_TIME TIME_ZONE COUNTY COUNTYNAME STATE
## 1     1.00  4/18/1950 0:00:00     0130       CST  97.00     MOBILE    AL
## 2     1.00  4/18/1950 0:00:00     0145       CST   3.00    BALDWIN    AL
## 3     1.00  2/20/1951 0:00:00     1600       CST  57.00    FAYETTE    AL
## 4     1.00   6/8/1951 0:00:00     0900       CST  89.00    MADISON    AL
## 5     1.00 11/15/1951 0:00:00     1500       CST  43.00    CULLMAN    AL
## 6     1.00 11/15/1951 0:00:00     2000       CST  77.00 LAUDERDALE    AL
##    EVTYPE BGN_RANGE BGN_AZI BGN_LOCATI END_DATE END_TIME COUNTY_END
## 1 TORNADO      0.00                                            0.00
## 2 TORNADO      0.00                                            0.00
## 3 TORNADO      0.00                                            0.00
## 4 TORNADO      0.00                                            0.00
## 5 TORNADO      0.00                                            0.00
## 6 TORNADO      0.00                                            0.00
##   COUNTYENDN END_RANGE END_AZI END_LOCATI LENGTH  WIDTH F  MAG FATALITIES
## 1         NA      0.00                   14.00 100.00 3 0.00       0.00
## 2         NA      0.00                    2.00 150.00 2 0.00       0.00
## 3         NA      0.00                    0.10 123.00 2 0.00       0.00
## 4         NA      0.00                    0.00 100.00 2 0.00       0.00
## 5         NA      0.00                    0.00 150.00 2 0.00       0.00
## 6         NA      0.00                    1.50 177.00 2 0.00       0.00
##   INJURIES PROPDMG PROPDMGEXP CROPDMG CROPDMGEXP WFO STATEOFFIC ZONENAMES
## 1    15.00   25.00          K    0.00
## 2     0.00    2.50          K    0.00
## 3     2.00   25.00          K    0.00
## 4     2.00    2.50          K    0.00
## 5     2.00    2.50          K    0.00
## 6     6.00    2.50          K    0.00
##   LATITUDE LONGITUDE LATITUDE_E LONGITUDE_ REMARKS REFNUM
## 1  3040.00   8812.00    3051.00    8806.00           1.00
## 2  3042.00   8755.00       0.00       0.00           2.00
## 3  3340.00   8742.00       0.00       0.00           3.00
## 4  3458.00   8626.00       0.00       0.00           4.00
## 5  3412.00   8642.00       0.00       0.00           5.00
## 6  3450.00   8748.00       0.00       0.00           6.00
```

Right away, we can see that this dataset will require a bit of cleaning to get into a more useful form. We see lots of missing entries, and further examination reveals that the 'EVTYPE' variable specifying the type of storm event contains 985 unique entries, many of which seem to be duplicate descriptions of the same/very similar event types (i.e. 'GUSTY THUNDERSTORM WINDS', 'GUSTY THUNDERSTORM WIND', 'TSTM WIND' are all listed as separate event types). We also have inconsistent entries in the 'PROPDMGEXP' and 'CROPDMGEXP' variables, which give a multiplier to get from the entries in 'PROPDMG' and 'CROPDMG' to actual dollar amounts of damage. For simplicity's sake, we'll start by removing all entries with zero damage or multipliers less than 'K/k' (1000x), as long as they also have no reported injuries or fatalities. This should declutter the data substantially by removing relatively minor events that wouldn't impact our overall assessment much. We'll also drop some additional variables that aren't relevant to our analysis, leaving only the following observations for each event: STATE, COUNTYNAME, EVTYPE, FATALITIES, INJURIES, PROPDMG/PROPDMGEXP, CROPDMG/CROPDMGEXP.

```
prelimclean <- subset(rawdata, (as.numeric(as.character(PROPDMG)) >
    0 & PROPDMGEXP %in% c("k", "K", "m", "M", "b", "B")) | (as.numeric(as.character(CROPDMG)) >
    0 & CROPDMGEXP %in% c("k", "K", "m", "M", "b", "B")) | as.numeric(as.character(INJURIES)) >
    0 | as.numeric(as.character(FATALITIES)) > 0, select = c(STATE,
    COUNTYNAME, EVTYPE, FATALITIES, INJURIES, PROPDMG, PROPDMGEXP,
    CROPDMG, CROPDMGEXP))
```

Now we have a dataset that is substantially smaller and contains only events with significant impacts. We will further clean up the data by doing a bit of dimension reduction. We can generate a total damage estimate by converting the PROPDMGEXP and CROPDMGEXP variables into numeric multipliers, and performing the necessary arithmetic (i.e. DMGTOTAL = PROPDMG*PROPDMGEXP + CROPDMG*CROPDMGEXP). We will also compress the INJURIES and FATALITIES down to a total CASUALTIES variable, with each injury counting for 3/10 the value of a fatality (in the absence of more data about the specific injuries, we want to preserve some sense of the magnitude of an event by weighting fatalities more heavily than injuries).

```
# Substitute the character multipliers with numeric ones
# given in the codebook using a lookup table In order to
# match the multiplier that's just an empty space, we must
# first replace the empty space with a character to match on
# for our substitution
prelimclean$PROPDMGEXP <- gsub("^$", "_", as.character(prelimclean$PROPDMGEXP))
prelimclean$CROPDMGEXP <- gsub("^$", "_", as.character(prelimclean$CROPDMGEXP))
lut <- c(`_` = 1, k = 1000, K = 1000, m = 1e+06, M = 1e+06, b = 1e+09,
    B = 1e+09)
reduceddata <- prelimclean %>% mutate(PROPDMGEXP = recode(PROPDMGEXP,
    !!!lut)) %>% mutate(CROPDMGEXP = recode(CROPDMGEXP, !!!lut))
# Replace the PROPDMG and CROPDMG fields with the calculated
# DMGTOTAL
reduceddata <- reduceddata %>% mutate(DMGTOTAL = (as.numeric(as.character(PROPDMG))) *
    PROPDMGEXP + (as.numeric(as.character(CROPDMG))) * CROPDMGEXP) %>%
    mutate(PROPDMG = NULL) %>% mutate(PROPDMGEXP = NULL) %>%
    mutate(CROPDMG = NULL) %>% mutate(CROPDMGEXP = NULL)
# Replace the INJURIES and FATALITIES fields with the
# calculated CASUALTIES
reduceddata <- reduceddata %>% mutate(CASUALTIES = (as.numeric(as.character(INJURIES))) *
    0.3 + (as.numeric(as.character(FATALITIES)))) %>% mutate(INJURIES = NULL) %>%
    mutate(FATALITIES = NULL)
```

Looking at the resulting table, we can see an event that looks like an outlier: flooding in Napa, CA resulting in $115 billion in damage. If we go back to the original data table and look up floods in Napa, we can see entries for a flood causing both $115 million and $115 billion in damages, so it looks like this is a data entry error (indeed, if we look up the corresponding event, widespread flooding in Northern California in winter 2005-2006, we see total damage estimates across several counties of about $300 million). We can simply remove the incorrect entry. The other very high-damage events seem to correspond well to actual historic incidents, so we can safely leave them alone. We'll also take this opportunity to remove any entries with 'NA' total damage (presumably resulting from unusual entries in either PROPDMGEXP or CROPDMGEXP in earlier steps), since there are only 37 such entries out of a total of about 250,000 so our analysis should be essentially unaffected.

```
maxID <- which.max(reduceddata$DMGTOTAL)
reduceddata <- reduceddata[-maxID, ]
reduceddata <- na.omit(reduceddata)
```

The last processing step we must perform is to condense the number of event types down from the current 985 to something a bit more manageable. We'll start by grouping our current set of events by event type and summing both the damage totals and casualty totals for each event type.

```
by_event_damage <- reduceddata %>% group_by(EVTYPE) %>% summarize(DMGTOTAL = sum(DMGTOTAL))
by_event_casualties <- reduceddata %>% group_by(EVTYPE) %>% summarize(CASUALTIES = sum(CASUALTIES))
```

Now we can sort in descending order of damage and casualties and examine the last few rows of each table to see if a small number of event types contribute the vast majority of the damage and/or casualty totals.

```r
by_event_damage <- by_event_damage %>% arrange(desc(DMGTOTAL))
head(by_event_damage, n = 15L)
```

```
## # A tibble: 15 x 2
##    EVTYPE               DMGTOTAL
##    <fct>                   <dbl>
##  1 HURRICANE/TYPHOON 71913712800
##  2 TORNADO           57301935590
##  3 STORM SURGE       43323541000
##  4 FLOOD             35287178250
##  5 HAIL              18733216176
##  6 FLASH FLOOD       17561538685
##  7 DROUGHT           15018672000
##  8 HURRICANE         14610229010
##  9 RIVER FLOOD       10148404500
## 10 ICE STORM          8967037810
## 11 TROPICAL STORM     8382236550
## 12 WINTER STORM       6715441250
## 13 HIGH WIND          5908617560
## 14 WILDFIRE           5060586800
## 15 TSTM WIND          5038935790
```

```r
by_event_casualties <- by_event_casualties %>% arrange(desc(CASUALTIES))
head(by_event_casualties, n = 15L)
```

```
## # A tibble: 15 x 2
##    EVTYPE            CASUALTIES
##    <fct>                  <dbl>
##  1 TORNADO               33026.
##  2 EXCESSIVE HEAT         3860.
##  3 TSTM WIND              2591.
##  4 FLOOD                  2507.
##  5 LIGHTNING              2385
##  6 HEAT                   1567
##  7 FLASH FLOOD            1511.
##  8 ICE STORM               682.
##  9 WINTER STORM            602.
## 10 HIGH WIND               587.
## 11 THUNDERSTORM WIND       579.
## 12 HURRICANE/TYPHOON       446.
## 13 RIP CURRENT             438.
## 14 HEAVY SNOW              430.
## 15 HAIL                    422.
```

It seems that tornadoes, thunderstorm wind, heat, flooding, and lightning account for most of the total casualties, and that hurricanes, tornadoes, flooding, drought, and hail cover most of the property/crop damage. We can also see some redundancies and duplications in the listed event types, and can combine them as follows:

- Hurricane:
  - Hurricane/Typhoon
  - Hurricane
  - Tropical Storm
  - Storm Surge
- WinterStorm:

    - Winter Storm
    - Ice Storm
    - Heavy Snow
    - Blizzard
- Thunderstorm:
    - TSTM Wind
    - Hail
    - Thunderstorm Wind
    - Thunderstorm Winds
    - High Wind
    - Lightning
- HeatEvent:
    - Excessive Heat
    - Heat
- Flooding:
    - Flood
    - Flash Flood
    - River Flood
- DroughtFire:
    - Drought
    - Wildfire

We can accomplish this consolidation by mapping the existing EVTYPE entries to our new condensed categories as follows:

```
evlookup <- c(`HURRICANE/TYPHOON` = "HURRICANE", HURRICANE = "HURRICANE",
    `TROPICAL STORM` = "HURRICANE", `STORM SURGE` = "HURRICANE",
    `STORM SURGE/TIDE` = "HURRICANE", `HURRICANE OPAL` = "HURRICANE",
    `WINTER STORM` = "WINTER STORM", `ICE STORM` = "WINTER STORM",
    `HEAVY SNOW` = "WINTER STORM", BLIZZARD = "WINTER STORM",
    `TSTM WIND` = "THUNDERSTORM", HAIL = "THUNDERSTORM", `THUNDERSTORM WIND` = "THUNDERSTORM",
    `THUNDERSTORM WINDS` = "THUNDERSTORM", `HIGH WIND` = "THUNDERSTORM",
    LIGHTNING = "THUNDERSTORM", `EXCESSIVE HEAT` = "HEATEVENT",
    HEAT = "HEATEVENT", FLOOD = "FLOODING", `FLASH FLOOD` = "FLOODING",
    `RIVER FLOOD` = "FLOODING", DROUGHT = "DROUGHTFIRE", WILDFIRE = "DROUGHTFIRE",
    TORNADO = "TORNADO", `RIP CURRENT` = "RIP CURRENT", `RIP CURRENTS` = "RIP CURRENT")
finaldata <- reduceddata %>% mutate(EVTYPE = recode(EVTYPE, !!!evlookup))
finaldata <- na.omit(finaldata)
final_by_damage <- finaldata %>% group_by(EVTYPE) %>% summarize(DMGTOTAL = sum(DMGTOTAL))
final_by_damage <- final_by_damage %>% arrange(desc(DMGTOTAL))
final_by_casualties <- finaldata %>% group_by(EVTYPE) %>% summarize(CASUALTIES = sum(CASUALTIES))
final_by_casualties <- final_by_casualties %>% arrange(desc(CASUALTIES))
```

This should get us to a reasonably compact list of event types to examine:
* Hurricane
* Winter Storm
* Thunderstorm
* Heat Event
* Flooding
* Drought/Fire
* Tornado
* Rip Current

We can calculate the percentage of total damage or casualties resulting from each event type and add that column to each table for later interpretation:

```
final_by_damage <- mutate(final_by_damage, DMGPCT = 100 * DMGTOTAL/sum(DMGTOTAL))
final_by_casualties <- mutate(final_by_casualties, CASUALTIESPCT = 100 *
    CASUALTIES/sum(CASUALTIES))
```

And finally, we will split off the top 6 event types by percentage of damage and casualties into separate tables for use in producing figures. We will also tweak some columns a bit for readability (changing column names to be more human-readable and converting our numbers down a few orders of magnitude to avoid a bunch of trailing zeroes).

```
topdamage <- final_by_damage[1:6, ]
topcasualties <- final_by_casualties[1:6, ]
topdamage <- topdamage %>% mutate(DMGTOTALBBN = DMGTOTAL/(10^9)) %>%
    mutate(DMGTOTAL = NULL)
colnames(topdamage) <- c("Event Type", "Percentage of Total Severe Weather Damage",
    "Total Damage (in billion USD)")
topcasualties <- topcasualties %>% mutate(KCASUALTIES = CASUALTIES/1000) %>%
    mutate(CASUALTIES = NULL)
colnames(topcasualties) <- c("Event Type", "Percentage of Total Severe Weather Casualties",
    "Total Casualties (thousands)")
```

Results

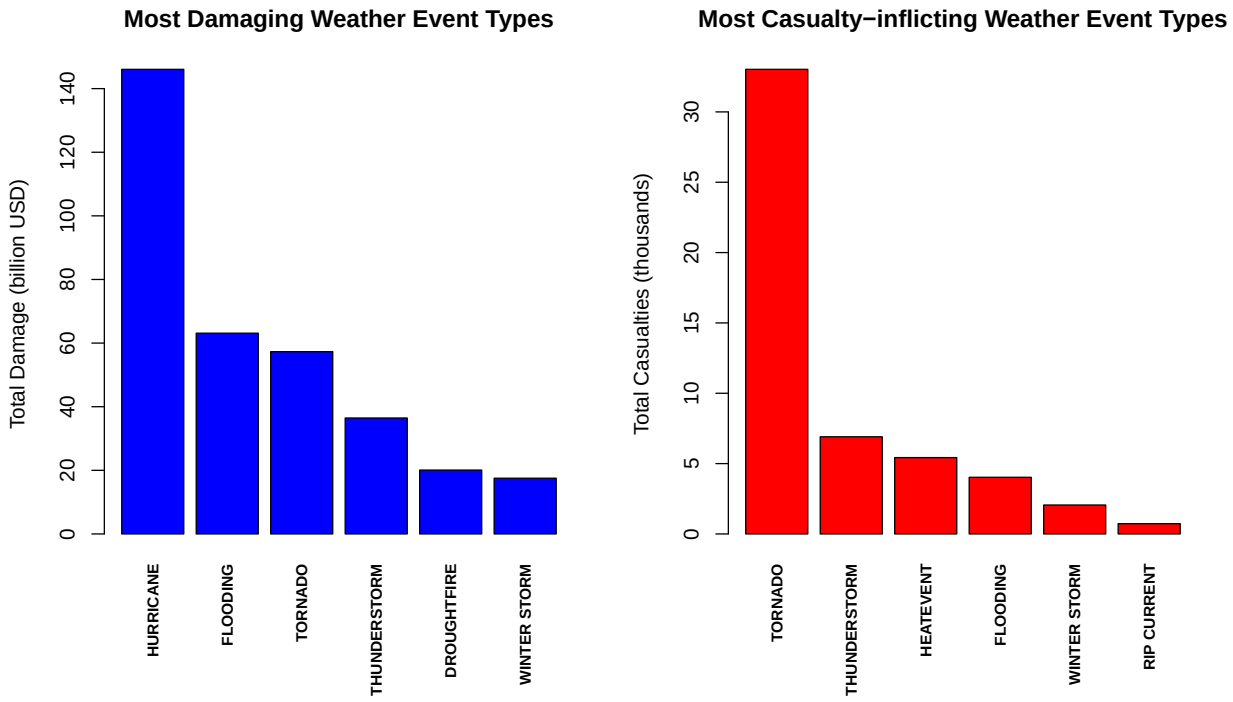
From our analysis, we identified 6 categories of severe weather events that were responsible for over 90% of total reported property/crop damage and casualties. A summary of these events and their associate damage or casualties follows in the tables below. We also present this data as a bar plot for a more intuitive look.

Table 1: Severe Weather Events by Total Damage

Event Type	Percentage of Total Severe Weather Damage	Total Damage (in billion USD)
HURRICANE	40.4	146
FLOODING	17.5	63
TORNADO	15.9	57
THUNDERSTORM	10.1	36
DROUGHTFIRE	5.6	20
WINTER STORM	4.8	18

Table 2: Severe Weather Events by Total Casualties

Event Type	Percentage of Total Severe Weather Casualties	Total Casualties (thousands)
TORNADO	57.7	33.03
THUNDERSTORM	12.1	6.91
HEATEVENT	9.5	5.43
FLOODING	7.0	4.03
WINTER STORM	3.6	2.06
RIP CURRENT	1.3	0.73



So we can clearly see that hurricanes, flood events, and tornadoes cause the most property and crop damage of the examined event types, while tornadoes cause the most casualties by far. Thunderstorm-related events cause both significant casualties and substantial property/crop damage. This makes a certain amount of sense: hurricanes and flooding are generally able to be predicted well in advance, allowing for evacuation or other measures to reduce casualties (but since buildings, agricultural operations, etc. can't be evacuated, substantial property damage still results). Tornadoes and thunderstorm impacts, by contrast, are significantly more localized and less predictable, making it difficult for preparations to be made to reduce casualties.