Reproducible Research: Peer Assessment 2

Connor Gooding

8/28/2017

Synopsis

The following analysis takes raw data from the National Weather Service regarding weather events in the United States. The first half of the report describes the process of transforming the raw data down to observations from 1996 and later across the 47 official weather event types listed by the National Weather Service. The second half of the report answers two questions:

1. Across the United States, which types of events are most harmful with respect to population health?
2. Across the United States, which types of events have the greatest economic consequences?

Data Processing

Getting the Raw Data

The raw data comes compressed in a .bz2 file, as it is quite large. The decompression and loading of this raw data into R takes some time, so this chunk of code is cached to save time.

```
theUrl <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
invisible(download.file(theUrl, "FStormData.csv.bz2"))
invisible(bunzip2('FStormData.csv.bz2', remove = FALSE, skip = TRUE))
FStormDataRaw <- as.data.frame(fread('FStormData.csv'))
```

Subsetting the Data

To make the cleaning process easier and more sensible, only observations from 1996 onward are going to be used in the analysis. This is because 1996 was the first year that the National Weather Service started to record all 48 of the current event types. To make the analysis more useful for municipal managers and state governors, only storms that occurred within the 50 US states will be considered.

Of the 37 variables presented in the raw data, only 8 of them will be useful in answering the questions posed for this analysis. They are: Year, Event Type, Fatalities, Injuries, Property Damage, Property Damage Exponent, Crop Damage, and Crop Damage Exponent (and the exponent columns will be eliminated shortly).

```
FStormData <- FStormDataRaw
data(state)
FStormData$BGN_DATE <- format(as.Date(sub(' [0-9]:[0-9][0-9]:[0-9][0-9]$',
                                    '', FStormData$BGN_DATE),
                                format = "%m/%d/%Y"), "%Y")
FStormData <- FStormData %>% filter(BGN_DATE >= 1996, STATE %in% state.abb) %>%
              select(2, 8, 23:28)
names(FStormData)[1] <- "YEAR"
```

Scaling Property and Crop Damages

In the raw data, the exponents for the scientific notation of property and crop damage values are isolated to their own columns. This prevents the observations from being compared against each other properly, and the

columns become redundant once used to operate on the raw damage value columns. The next chunk of code does said operation and then eliminates these columns.

```
## adjust the damage exponents
source('FStormExponents.R', local = T)
FStormData$PROPDMG <- as.numeric(mapply(FStormExponents, FStormData$PROPDMG,
                                FStormData$PROPDMGEXP, SIMPLIFY = T))
FStormData$CROPDMG <- as.numeric(mapply(FStormExponents, FStormData$CROPDMG,
                                FStormData$CROPDMGEXP, SIMPLIFY = T))

FStormData <- FStormData %>% select(YEAR, EVTYPE, FATALITIES,
                                    INJURIES, PROPDMG, CROPDMG)
```

Wrangling Event Types

Perhaps the most problematic column in the raw data is EVTYPE, which is supposed to classify each observation as one of the 48 official storm event types as documented by the National Weather Service. However, due to typos and other differences in documentation, there are more than 48 unique values in the raw data.

```
length(unique(FStormData$EVTYPE))
```

```
## [1] 495
```

The solution to this is to simply do the best we can to wrangle these erroneous event types into the set of valid event types and eliminate those that cannot be classified. This is accomplished by initially running through a series of regular expression calls, then approximately matching the remaining offenders to the set of event types.

To accomplish this portion of the data transformation process, an additional .csv file was included into the project. It can be found here

```
## Load the list of official event types (I created this and included it in the repo).
EventTypes <- read.csv('EventTypes.csv')

## Cast all event types to all-caps.
FStormData$EVTYPE <- toupper(FStormData$EVTYPE)

## Wrangle thunderstorm winds and marine thunderstorm winds.
FStormData$EVTYPE <- sub('TSTM', 'THUNDERSTORM',
                         FStormData$EVTYPE)
FStormData$EVTYPE[grep('MARINE THUNDER', FStormData$EVTYPE)] <- 'MARINE TSTM WIND'
FStormData$EVTYPE[grep('THUNDERSTORM WIND', FStormData$EVTYPE)] <- 'THUNDERSTORM WIND'

## Wrangle commonly-reported but outdated event types
FStormData$EVTYPE[grep('FLASH', FStormData$EVTYPE)] <- 'FLASH FLOOD'
FStormData$EVTYPE[grep('URBAN', FStormData$EVTYPE)] <- 'HEAVY RAIN'
FStormData$EVTYPE[grep('FIRE', FStormData$EVTYPE)] <- 'WILDFIRE'
FStormData$EVTYPE[grep('^WIN(.*)MIX$', FStormData$EVTYPE)] <- 'WINTER WEATHER'
FStormData$EVTYPE[grep('EXTREME', FStormData$EVTYPE)] <- 'EXTREME COLD/WIND CHILL'
FStormData$EVTYPE <- sub('LANDSLIDE', 'DEBRIS FLOW', FStormData$EVTYPE)
FStormData$EVTYPE <- sub('^FOG$', 'DENSE FOG', FStormData$EVTYPE)
FStormData$EVTYPE <- sub('^SNOW$', 'HEAVY SNOW', FStormData$EVTYPE)
FStormData$EVTYPE <- sub('^WIND$', 'STRONG WIND', FStormData$EVTYPE)
FStormData$EVTYPE[grep('SURGE', FStormData$EVTYPE)] <- 'STORM TIDE'
FStormData$EVTYPE[grep('SURF', FStormData$EVTYPE)] <- 'HIGH SURF'
```

```
FStormData$EVTYPE[grep('HURRICANE', FStormData$EVTYPE)] <- 'HURRICANE/TYPHOON'

## Approximately match the rest to the nearest event type
FStormData$EVTYPE <- EventTypes$x[amatch(FStormData$EVTYPE, EventTypes$x, maxDist = 5)]
```

The full process wrangles the observations to the nearest valid event type with a success rate of 99.8%.

```
percent((1 - (sum(is.na(FStormData$EVTYPE)) / length(FStormData$EVTYPE))))
```

```
## [1] "99.8%"
```

The .2% portion of observations that did not survive the event casting will be discarded for the analyses to come.

```
FStormDataClean <- FStormData %>% filter(!(is.na(EVTYPE)))
```

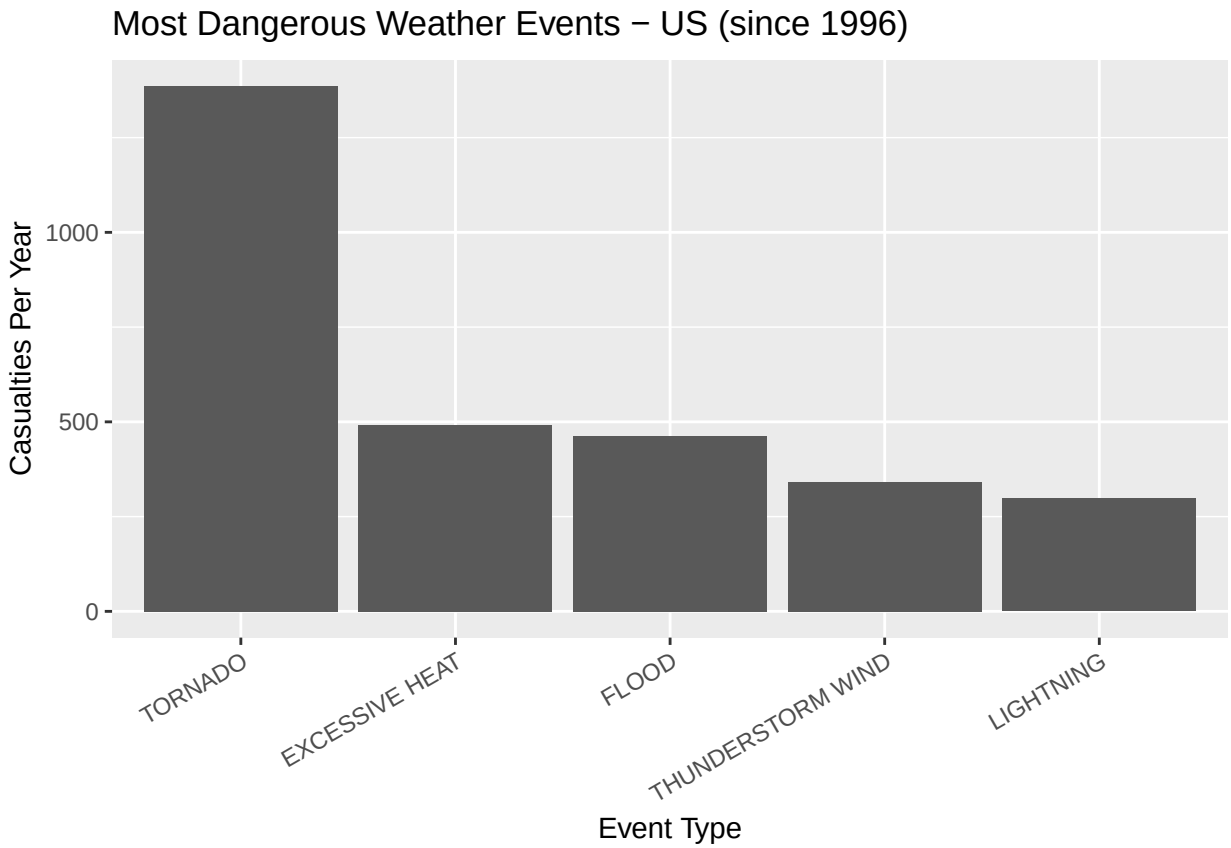
Results

Across the United States, which types of events are most harmful with respect to population health?

The most harmful events with respect to population health will be defined as the ones that have the highest yearly casualty rates. Summing together fatalities and injuries and finding the mean across all years since 1996, the following barplot lists the five events with the highest yearly casualty rates.

```
Casualties <- FStormDataClean %>% group_by(EVTYPE, YEAR) %>%
    summarize(Casualties = sum(INJURIES + FATALITIES)) %>% group_by(EVTYPE) %>%
    summarize(CasualtyRate = mean(Casualties)) %>% arrange(desc(CasualtyRate))

ggplot(data = head(Casualties, 5),
       aes(x = reorder(EVTYPE, -CasualtyRate), y = CasualtyRate)) +
       geom_bar(stat = "identity") + xlab("Event Type") +
       ylab("Casualties Per Year") +
       ggtitle("Most Dangerous Weather Events - US (since 1996)") +
       theme(axis.text.x = element_text(angle=30, hjust=1))
```



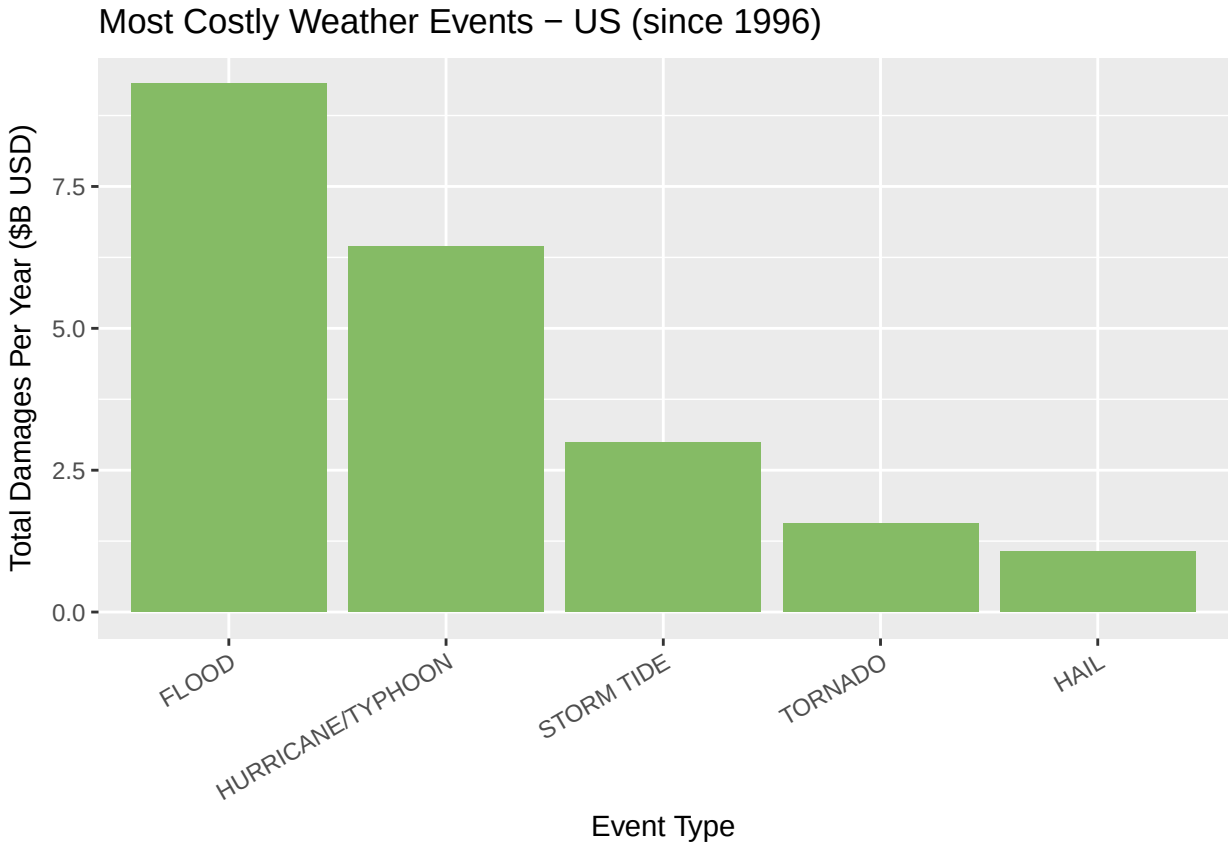
From the barplot, it appears that tornadoes are definitively responsible for the highest yearly casualty rate of all storms in the 50 US states since 1996 with almost 1500 casualties a year. Following farther behind, excessive heat, floods, thunderstorm winds, and lightning are the next most dangerous storm events.

Across the United States, which types of events have the greatest economic consequences?

The most economically disatrous weather events will be the ones that do the most yearly damage to crops and/or property. Summing together crop and propety damages and finding the mean across all years since 1996, the following barplot lists the five events with the highest yearly damage costs.

```
AllDamages <- FStormDataClean %>% group_by(EVTYPE, YEAR) %>%
    summarize(TotalDamages = sum(CROPDMG + PROPDMG)) %>% group_by(EVTYPE) %>%
    summarize(DamageRate = mean(TotalDamages)) %>% arrange(desc(DamageRate))

ggplot(data = head(AllDamages, 5),
       aes(x = reorder(EVTYPE, -DamageRate), y = DamageRate/1e9)) +
       geom_bar(stat = "identity", fill = "#85bb65") + xlab("Event Type") +
       ylab("Total Damages Per Year ($B USD)") +
       ggtitle("Most Costly Weather Events - US (since 1996)") +
       theme(axis.text.x = element_text(angle=30, hjust=1))
```



From this barplot, it is evident that flooding is responsible for the most economic damage yearly, with almost $10B in combined crop and property damage per year. Hurricanes and typhoons come in second with just over $6.25B, with storm tides, tornadoes, and hail rounding out the top five.

Conclusions

Based on the first barplot, tornadoes have been the most hazardous weather events to public health in the United States over the past two decades. Tornadoes have been responsible for almost 1500 annual casualties since 1996, more than double that of the next most dangerous weather event, excessive heat. Following tornadoes and excessive heat, the next three most dangerous weather events were floods, thunderstorm winds, and lightning.

Based on the second barplot, floods have been the most economically costly weather events in the United States over the past two decades. Floods have been responsible for over 9 billion dollars of damage to crops and property per year, eclipsing the combined economic effect of hurricanes and typhoons per year. Following floods and hurricanes/typhoons, the next three most costly weather events annually were storm tides, tornadoes, and hail.