Natural Disaster in USA (1950-2011): Human and Economic Consequences

Aohagi

October 6, 2018

Synopsis

Storms and other severe weather events can cause both public health and economic problems for communities and municipalities. Many severe events can result in fatalities, injuries, and property damage, and preventing such outcomes to the extent possible is a key concern.

This project involves exploring the U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database. This database tracks characteristics of major storms and weather events in the United States, including when and where they occur, as well as estimates of any fatalities, injuries, and property damage.

The data for this assignment come in the form of a comma-separated-value file compressed via the bzip2 algorithm to reduce its size. You can download the file from the course web site: Storm Data (47Mb) There is also some documentation of the database available. Here you will find how some of the variables are constructed/defined. National Weather Service Storm Data Documentation and the most frequently asked question here National Climatic Data Center Storm Events FAQ

The events in the database start in the year 1950 and end in November 2011. In the earlier years of the database there are generally fewer events recorded, most likely due to a lack of good records. More recent years should be considered more complete.

Getting the data

I first created a project folder if it not already exist and dowloaded the data from the website to the project folder.

```
# creating project folder and getting the data
if (!file.exists('Project 2')) {
        dir.create('Project 2')
}
# I have already downloaded the data, so I'm commiting out but if you need you can just delete the '#'
#download.file("https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2", "Project 2/s
```

Data Processing

Here, I loaded the zipped data with the readr package and saved it to the object storm.

```
# attaching the reader package and loading the storm dat
library(readr)
storm <- read_csv("storm.csv.bz2")
```

This project tries to answer these two questions:

1. Across the United States, which types of events (as indicated in the EVTYPE variable) are most harmful with respect to population health?
2. Across the United States, which types of events have the greatest economic consequences?

It's important to remember that the data should be processed accoriding to the questions it's trying to answer. It seems that we only need a subset of this dataset, so i will just select the variable needed for our analysis: 'EVTYPE', 'FATALITIES', 'INJURIES', 'PROPDMG', 'PROPDMGEXP', 'CROPDMG', 'CROPDMGEXP'.

```r
# loading dplyr package and subsetting the data
library(dplyr)
storm_sub <- select(storm, c('EVTYPE', 'FATALITIES', 'INJURIES', 'PROPDMG', 'PROPDMGEXP', 'CROPDMG', 'C
```

inspecting the subset data to see what the variable are and how they are recorded.

```r
# showing the first view rows
head(storm_sub)
```

```
# A tibble: 6 x 7
  EVTYPE  FATALITIES INJURIES PROPDMG PROPDMGEXP CROPDMG CROPDMGEXP
  <chr>        <dbl>    <dbl>   <dbl> <chr>        <dbl> <chr>
1 TORNADO          0       15    25   K                0 <NA>
2 TORNADO          0        0     2.5 K                0 <NA>
3 TORNADO          0        2    25   K                0 <NA>
4 TORNADO          0        2     2.5 K                0 <NA>
5 TORNADO          0        2     2.5 K                0 <NA>
6 TORNADO          0        6     2.5 K                0 <NA>
```

We see that we can answer our first question with the first 3 variables (i.e. 'EVTYPE', 'FATALITIES', 'INJURIES') and they are nicely recorded.

But it seems that the rest variables need some transformation to answer our second question. Notice that the 'EXP' of the PROPDMGEXP and CROPDMGEXP variables recorded as characters means *exponent with base 10* to the PROPDMG and CROPDMG variables respectively.

It turned out that some processing is necessary before analysis. First, I print out the unique values of the PROPDMGEXP and CROPDMGEXP and then specify what they were designated for.

```r
# showing the unique values of PROPDMGEXP and CROPDMGEXP
unique(storm_sub$PROPDMGEXP)
```

```
 [1] "K" "M" NA  "B" "m" "+" "0" "5" "6" "?" "4" "2" "3" "h" "7" "H" "-"
[18] "1" "8"
```

```r
unique(storm_sub$CROPDMGEXP)
```

```
[1] NA  "M" "K" "m" "B" "?" "0" "k" "2"
```

I believe the values are coded as follows:

"H" = 100, "h" = 100, "K" = 1000, "k" = 1000, "M" = 1e06, "m" = 1e06, "B" = 1e09, "m" = 1e06, "8" = 1e08, "7" = 1e07, "6" = 1e06, "5" = 1e05, "4" = 1e04, "3" = 1e03, "2" = 1e02, "1" = 10 , "0" = 0, "-" = 0, "+" = 0, "?" = 0

Finally, to answer the second question, we need to create two new variables: one for propery damage and the other for crop damage. The two new variables, **PROPDMGVAL & CROPDMGVAL**, are damages valued in $ dollars as products of **PROPDMG & PROPDMGEXP** and **CROPDMG & CROPDMGEXP** respectively.

```r
# changing the PROPDMGEXP values to the numbers they designate
storm_sub$PROPDMGEXP[storm_sub$PROPDMGEXP == "K"] <- 1e03
storm_sub$PROPDMGEXP[storm_sub$PROPDMGEXP == "M"] <- 1e06
storm_sub$PROPDMGEXP[storm_sub$PROPDMGEXP == "B"] <- 1e09
storm_sub$PROPDMGEXP[storm_sub$PROPDMGEXP == "m"] <- 1e06
storm_sub$PROPDMGEXP[storm_sub$PROPDMGEXP == "H"] <- 1e02
storm_sub$PROPDMGEXP[storm_sub$PROPDMGEXP == "h"] <- 1e02
storm_sub$PROPDMGEXP[storm_sub$PROPDMGEXP == "+"] <- 0
storm_sub$PROPDMGEXP[storm_sub$PROPDMGEXP == "-"] <- 0
storm_sub$PROPDMGEXP[storm_sub$PROPDMGEXP == "?"] <- 0
```

```
storm_sub$PROPDMGEXP[storm_sub$PROPDMGEXP == "0"] <- 0
storm_sub$PROPDMGEXP[storm_sub$PROPDMGEXP == "1"] <- 1e01
storm_sub$PROPDMGEXP[storm_sub$PROPDMGEXP == "2"] <- 1e02
storm_sub$PROPDMGEXP[storm_sub$PROPDMGEXP == "3"] <- 1e03
storm_sub$PROPDMGEXP[storm_sub$PROPDMGEXP == "4"] <- 1e04
storm_sub$PROPDMGEXP[storm_sub$PROPDMGEXP == "5"] <- 1e05
storm_sub$PROPDMGEXP[storm_sub$PROPDMGEXP == "6"] <- 1e06
storm_sub$PROPDMGEXP[storm_sub$PROPDMGEXP == "7"] <- 1e07
storm_sub$PROPDMGEXP[storm_sub$PROPDMGEXP == "8"] <- 1e08

# transforming the PROPDMGEXP to numeric
storm_sub$PROPDMGEXP <- as.numeric(storm_sub$PROPDMGEXP)

# creating the PROPDMGVAL as the product of the other two
storm_sub <- mutate(storm_sub, PROPDMGVAL = storm_sub$PROPDMG * storm_sub$PROPDMGEXP)
```

```
# changing the CROPDMGEXP values to the numbers they designate
storm_sub$CROPDMGEXP[storm_sub$CROPDMGEXP == "K"] <- 1e03
storm_sub$CROPDMGEXP[storm_sub$CROPDMGEXP == "M"] <- 1e06
storm_sub$CROPDMGEXP[storm_sub$CROPDMGEXP == "m"] <- 1e06
storm_sub$CROPDMGEXP[storm_sub$CROPDMGEXP == "B"] <- 1e09
storm_sub$CROPDMGEXP[storm_sub$CROPDMGEXP == "?"] <- 0
storm_sub$CROPDMGEXP[storm_sub$CROPDMGEXP == "k"] <- 1e03
storm_sub$CROPDMGEXP[storm_sub$CROPDMGEXP == "2"] <- 1e02
storm_sub$CROPDMGEXP[storm_sub$CROPDMGEXP == "0"] <- 0

# transforming the CROPDMGEXP to numeric
storm_sub$CROPDMGEXP <- as.numeric(storm_sub$CROPDMGEXP)

# creating the CROPDMGEXP as the product of the other two
storm_sub <- mutate(storm_sub, CROPDMGVAL = storm_sub$CROPDMG * storm_sub$CROPDMGEXP)
```

The last step of preparing the data for analysis is to the select the 5 variables that will answer the two questions of the project. These are 'EVTYPE', 'FATALITIES' 'INJURIES', 'PROPDMGVAL' and 'CROPDMGVAL'.

```
# subsetting the data by select the 5 variables
storm_sub5 <- storm_sub %>% select('EVTYPE', 'FATALITIES', 'INJURIES', 'PROPDMGVAL', 'CROPDMGVAL')
```

Results

In this part of the project, you will find tables, plots and summary statistics of subsets of the dataset.

First Question

When we answering the most harmful events with respect to population health, then the variables we consider shoud be the number of injuries and fatalities of each event.

```
# grouping the data by event and summarizing the injuries and fatalities totals by each event
storm_health <- storm_sub5 %>% select('EVTYPE', 'FATALITIES', 'INJURIES') %>% group_by(EVTYPE) %>%
                summarize(Fatalities = sum(FATALITIES) ,Injuries = sum(INJURIES)) %>% arrange(desc(Fatal
storm_health
```

```
# A tibble: 977 x 3
   EVTYPE         Fatalities Injuries
   <chr>               <dbl>    <dbl>
 1 TORNADO              5633    91346
```

```
 2 EXCESSIVE HEAT       1903     6525
 3 FLASH FLOOD           978     1777
 4 HEAT                  937     2100
 5 LIGHTNING             816     5230
 6 TSTM WIND             504     6957
 7 FLOOD                 470     6789
 8 RIP CURRENT           368      232
 9 HIGH WIND             248     1137
10 AVALANCHE             224      170
# ... with 967 more rows
```

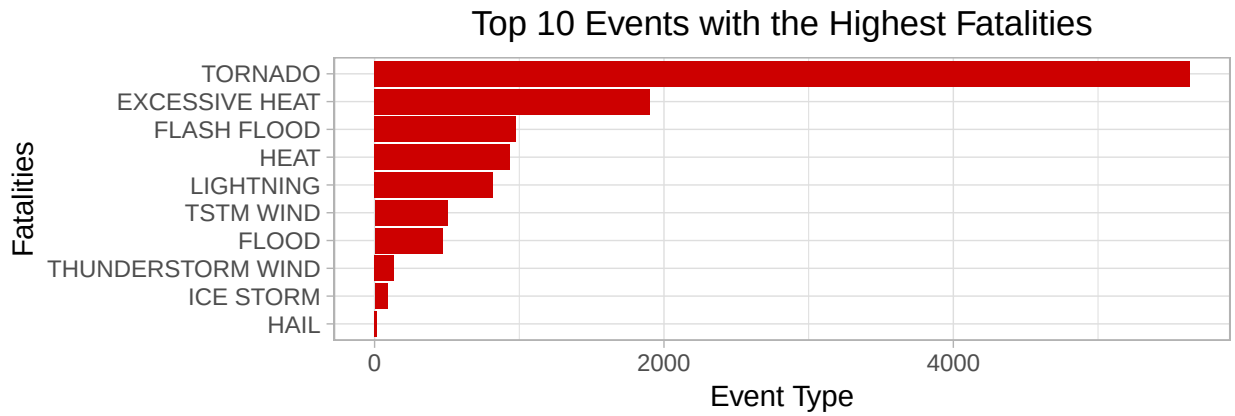
We see above the events with the highest fatalities and the corresponding injuries in each event.

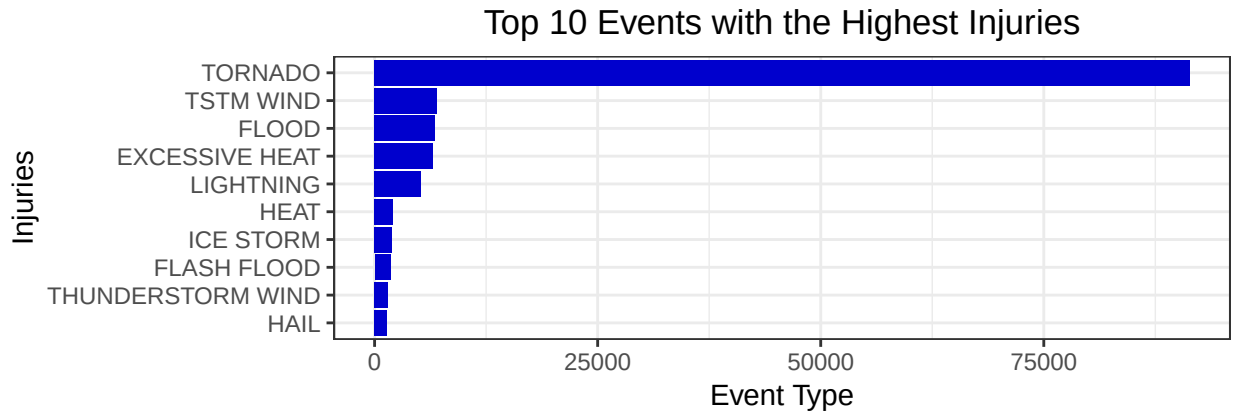
```
# group the data by event and summarizing the total of injuries and fatalities by each event
storm_health <- storm_sub5 %>% select('EVTYPE', 'FATALITIES', 'INJURIES') %>% group_by(EVTYPE) %>%
                        summarize(Fatalities = sum(FATALITIES) ,Injuries = sum(INJURIES)) %>% summary()
storm_health
```

```
    EVTYPE            Fatalities        Injuries
 Length:977         Min.   :   0.0   Min.   :    0.0
 Class :character   1st Qu.:   0.0   1st Qu.:    0.0
 Mode  :character   Median :   0.0   Median :    0.0
                    Mean   :  15.5   Mean   :  143.8
                    3rd Qu.:   0.0   3rd Qu.:    0.0
                    Max.   :5633.0   Max.   :91346.0
```

There are 977 different events and the mean of injuries and fatalities by event is 15.5 and 143.8 respectively but median, and the 75th percentile of both injuries and fatalities distribution are zero. However, there are still some very harmful events with injuries and fatalities as high as 91346 and 5633 respectively.

```
# loading ggplot2
library(ggplot2)

# top 10 events with highest human casualties
top10 <- storm_sub5 %>% select('EVTYPE', 'FATALITIES', 'INJURIES') %>% group_by(EVTYPE) %>%
                        summarize(Fatalities = sum(FATALITIES) ,Injuries = sum(INJURIES)) %>% top_n(10)

# top 10 fatalities by Event Type
g1 <- ggplot(top10, aes(reorder(x = EVTYPE, Fatalities), y = Fatalities)) + geom_bar(stat = "identity",

# top 10 injuries by Event Type
g2 <- ggplot(top10, aes(reorder(x = EVTYPE, Injuries), y = Injuries)) + geom_bar(stat = "identity", fil

# ploting both on the same plot
require(gridExtra)
grid.arrange(g1, g2)
```





This plot shows the top 10 events with the highest fatalities and injuries across USA over time. In both injuries and fatalities, we can see that **Tornado** is the most harmful disaster to human health.

Second Question

Now, answering the types of events that have the greatest economic consequences, we use our newly created variables: PROPDMGVAL & CROPDMGVAL of each event

```
# grouping the data by event and summarizing the PROPDMGVAL and CROPDMGVAL totals by each event
storm_econ <- storm_sub5 %>% select('EVTYPE', 'PROPDMGVAL', 'CROPDMGVAL') %>% group_by(EVTYPE) %>%
 summarize(PROPDMGVAL = sum(PROPDMGVAL, na.rm = T)/1e9 ,CROPDMGVAL = sum(CROPDMGVAL, na.rm = T)/1e9) %>%
 arrange(desc(PROPDMGVAL))
storm_econ
```

```
# A tibble: 977 x 3
   EVTYPE            PROPDMGVAL CROPDMGVAL
   <chr>                  <dbl>      <dbl>
 1 FLOOD                145.      5.66
 2 HURRICANE/TYPHOON     69.3     2.61
 3 TORNADO               56.9     0.415
 4 STORM SURGE           43.3     0.000005
 5 FLASH FLOOD           16.8     1.42
 6 HAIL                  15.7     3.03
 7 HURRICANE             11.9     2.74
 8 TROPICAL STORM         7.70    0.678
 9 WINTER STORM           6.69    0.0269
10 HIGH WIND              5.27    0.639
```

```
# ... with 967 more rows
```

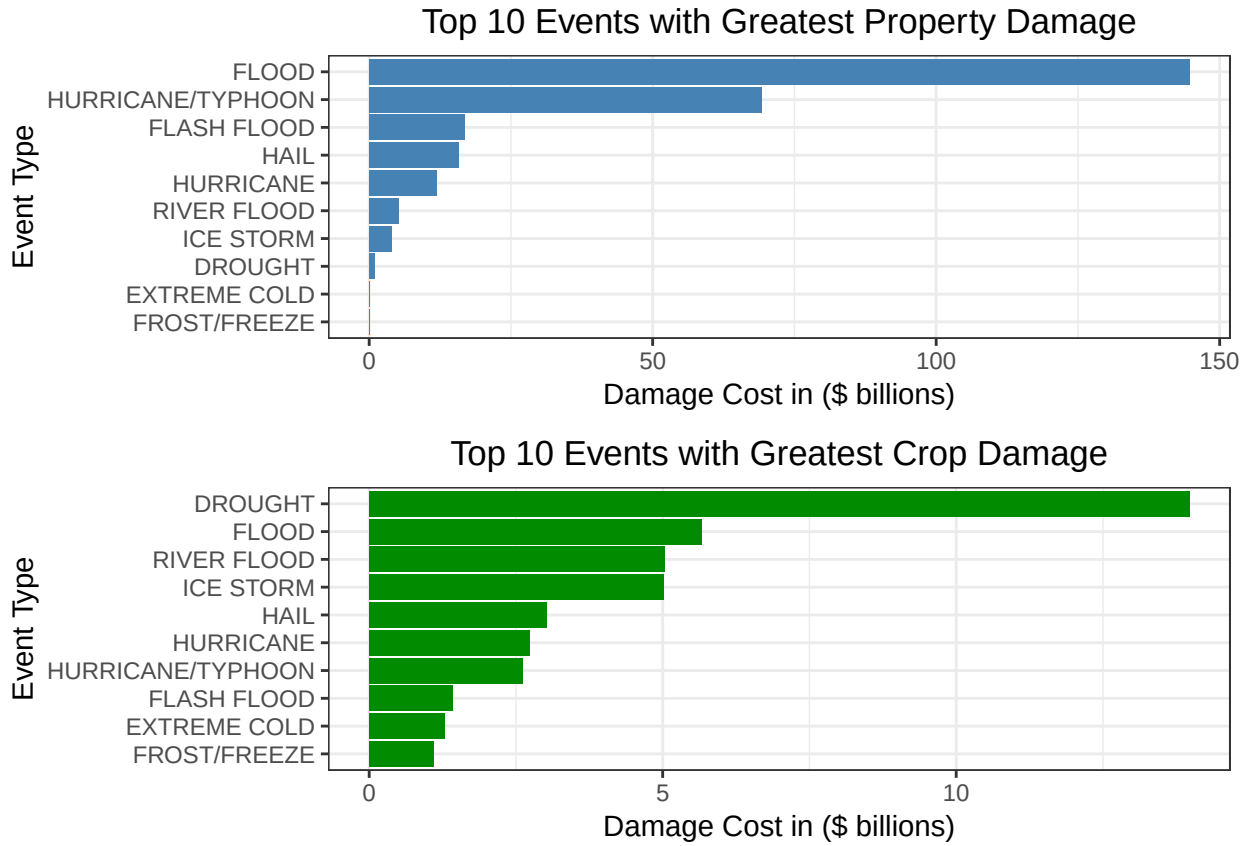
These are the events with the greatest economic consequences ordered by PROPDMGVAL in billions of US dollars.

```
# group the data by event and summarizing the total of PROPDMGVAL and CROPDMGVAL by each event
storm_econ <- storm_sub5 %>% select('EVTYPE', 'PROPDMGVAL', 'CROPDMGVAL') %>% group_by(EVTYPE) %>%
summarize(PROPDMGVAL = sum(PROPDMGVAL, na.rm = T)/1e9 ,CROPDMGVAL = sum(CROPDMGVAL, na.rm = T)/1e9) %>%
storm_econ
```

```
    EVTYPE            PROPDMGVAL          CROPDMGVAL
 Length:977         Min.   :  0.00000   Min.   : 0.00000
 Class :character   1st Qu.:  0.00000   1st Qu.: 0.00000
 Mode  :character   Median :  0.00000   Median : 0.00000
                    Mean   :  0.43831   Mean   : 0.05026
                    3rd Qu.:  0.00005   3rd Qu.: 0.00000
                    Max.   :144.65771   Max.   :13.97257
```

This summary is in billions of US dollars. There is economic consequences in property damage and crop damage as high as 144.658 and 13.973 in billions of dollars respectively.

```
# top 10 events with highest greates economic consequences
top10 <- storm_sub5 %>% select('EVTYPE', 'PROPDMGVAL', 'CROPDMGVAL') %>% group_by(EVTYPE) %>%
summarize(PROPDMGVAL = sum(PROPDMGVAL, na.rm = T)/1e9 ,CROPDMGVAL = sum(CROPDMGVAL, na.rm = T)/1e9) %>%

# top 10 property damage by Event Type
g3 <- ggplot(top10, aes(reorder(x = EVTYPE, PROPDMGVAL), y = PROPDMGVAL)) + geom_bar(stat = "identity",

# top 10 crop damage by Event Type
g4 <- ggplot(top10, aes(reorder(x = EVTYPE, CROPDMGVAL), y = CROPDMGVAL)) + geom_bar(stat = "identity",

# ploting both on the same plot
require(gridExtra)
grid.arrange(g3, g4)
```



When it comes to property damage, **Flood** stands out with 144.658 billions worth damages. It's worth mentioning that **Hurricane/Typhoon** does a considerable amount of damages. Not surprisingly, **Drought** causes the greatest crop damage worth 13.973 billions of dollars. In addition, **Flood**, **River Flood** and **Ice Storm** contribute a noticeable share to crop damages