Impact of Severe Weather Events on US Public Health and Economy

Riccardo Finotello

18 June 2020

Synopsis

Severe weather conditions and events can have a deep impact on public health with an increase in fatalities and injuries as well as on economy, causing temporary and permanent damage. With this analysis we explore the NOAA Storm Database containing data on atmospheric events from 1950 to 2011 in the US and we try to determine the aspects most harmful to population health and economy.

Data Processing

The focus of this section is to download, load and prepare the data for the analysis. We first download the file, available at this URL:

```
file.url <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
file.out <- "data.csv.bz2"
if(!file.exists(file.out)) {download.file(file.url, file.out, method = "curl")}
```

We read the full file using `read.csv` (which can read *bz2* compressed files) and then we convert it to a `data.table`:

```
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 3.6.3
```

```
data <- read.csv(file.out, header = TRUE, stringsAsFactors = FALSE)
data <- data.table(data)
```

First of all we notice the dimensions of the dataset, since it will directly influence processing time for every transformation we will perform:

```
print(paste("No. of samples:", dim(data)[1]))
```

```
## [1] "No. of samples: 902297"
```

```
print(paste("No. of columns:", dim(data)[2]))
```

```
## [1] "No. of columns: 37"
```

The dataset is therefore made of 37 columns named:

```
colnames(data)
```

```
##  [1] "STATE__"    "BGN_DATE"   "BGN_TIME"   "TIME_ZONE"  "COUNTY"
##  [6] "COUNTYNAME" "STATE"      "EVTYPE"     "BGN_RANGE"  "BGN_AZI"
## [11] "BGN_LOCATI" "END_DATE"   "END_TIME"   "COUNTY_END" "COUNTYENDN"
## [16] "END_RANGE"  "END_AZI"    "END_LOCATI" "LENGTH"     "WIDTH"
## [21] "F"          "MAG"        "FATALITIES" "INJURIES"   "PROPDMG"
```

```
## [26] "PROPDMGEXP" "CROPDMG"    "CROPDMGEXP" "WFO"        "STATEOFFIC"
## [31] "ZONENAMES"  "LATITUDE"   "LONGITUDE"  "LATITUDE_E" "LONGITUDE_"
## [36] "REMARKS"    "REFNUM"
```

whose classes are:

```
col.class <- sapply(data, class)
```

In this analysis we wil be particularly focused on **public health** related consequences and **economic** damage per event type. In order to speed up some computations we restrict the data we use to the **date** of the occurrence (the column `BGN_DATE`), the **event type** (`EVTYPE`), fatalities and injuries (`FATALITIES` and `INJURIES` respectively), property and crop damage (`PROPDMG` and `CROPDMG`, each expressed in **billion** of US$):

```
data.analysis <- data[, c("BGN_DATE",
                          "EVTYPE",
                          "FATALITIES",
                          "INJURIES",
                          "PROPDMG",
                          "CROPDMG"
                         )
                     ]
```

We then transform the date column into a `Date` class and rename the columns:

```
data.analysis[, BGN_DATE := as.Date(BGN_DATE, "%m/%d/%Y %H:%M:%S"),]
colnames(data.analysis) <- make.names(c("Date",
                                        "Type",
                                        "Fatalities",
                                        "Injuries",
                                        "Property.damage",
                                        "Crop.damage"
                                       )
                                      )
```

The new dataset has now 6 columns whose classes are:

```
col.class.analysis <- sapply(data.analysis, class)
col.class.analysis
```

```
##            Date            Type      Fatalities        Injuries Property.damage
##          "Date"     "character"       "numeric"       "numeric"       "numeric"
##     Crop.damage
##       "numeric"
```

The tidied database can now be used for the analysis. We first check the presence of missing values:

```
for(column in colnames(data.analysis)) {
    na.val = sum(as.numeric(is.na(column)))
    print(paste("Missing values in ", column, ": ", na.val, sep = ""))
}
```

```
## [1] "Missing values in Date: 0"
## [1] "Missing values in Type: 0"
## [1] "Missing values in Fatalities: 0"
## [1] "Missing values in Injuries: 0"
## [1] "Missing values in Property.damage: 0"
## [1] "Missing values in Crop.damage: 0"
```

We can therefore provide a summary of the dataset without worrying about any strategy to replace missing values:

```
summary(data.analysis)
```

```
##       Date                Type             Fatalities          Injuries
##  Min.   :1950-01-03   Length:902297      Min.   :  0.0000   Min.   :   0.0000
##  1st Qu.:1995-04-20   Class :character   1st Qu.:  0.0000   1st Qu.:   0.0000
##  Median :2002-03-18   Mode  :character   Median :  0.0000   Median :   0.0000
##  Mean   :1998-12-27                      Mean   :  0.0168   Mean   :   0.1557
##  3rd Qu.:2007-07-28                      3rd Qu.:  0.0000   3rd Qu.:   0.0000
##  Max.   :2011-11-30                      Max.   :583.0000   Max.   :1700.0000
##  Property.damage    Crop.damage
##  Min.   :   0.00   Min.   :  0.000
##  1st Qu.:   0.00   1st Qu.:  0.000
##  Median :   0.00   Median :  0.000
##  Mean   :  12.06   Mean   :  1.527
##  3rd Qu.:   0.50   3rd Qu.:  0.000
##  Max.   :5000.00   Max.   :990.000
```

We also provide the plot of the collected data per year to establish the significance of the study:

```
library(ggplot2)
n.events <- data.analysis[, .N, by = year(Date)] # count the events per year
g <- ggplot(data = n.events, aes(x = year, y = N))
g + geom_bar(stat = "identity") +
    xlab("year") +
    ylab("reported events") +
    ggtitle("Time Evolution of Reported Atmospheric Events in the US")
```

We see that the number of reported events has grown in time most probably due to more rigorous records and time series. This will clearly affect the results of the analysis.

Results

This section is focused on results. We provide a panoramic view on what had the largest impact on public health and economy using the previously introduced dataset.

Outcome on Public Health

From the data, we can recover the total amount of fatalities and injuries grouped by the type of atmospheric event:

```
data.type <- data.analysis[, .(Total.fatalities = sum(Fatalities),
                               Total.injuries   = sum(Injuries)
                              ),
                           by = Type
                          ]

# other than the total amount, we add the percentage of fatalities and injuries
data.type[, Perc.fatalities := Total.fatalities / sum(Total.fatalities)]
data.type[, Perc.injuries   := Total.injuries / sum(Total.injuries)]
```

We can then investigate the 10 most influencial causes of deaths and damage to people:

```
# order by percentage
data.fatalities <- data.type[order(-data.type$Perc.fatalities),]
data.injuries   <- data.type[order(-data.type$Perc.injuries),]
```

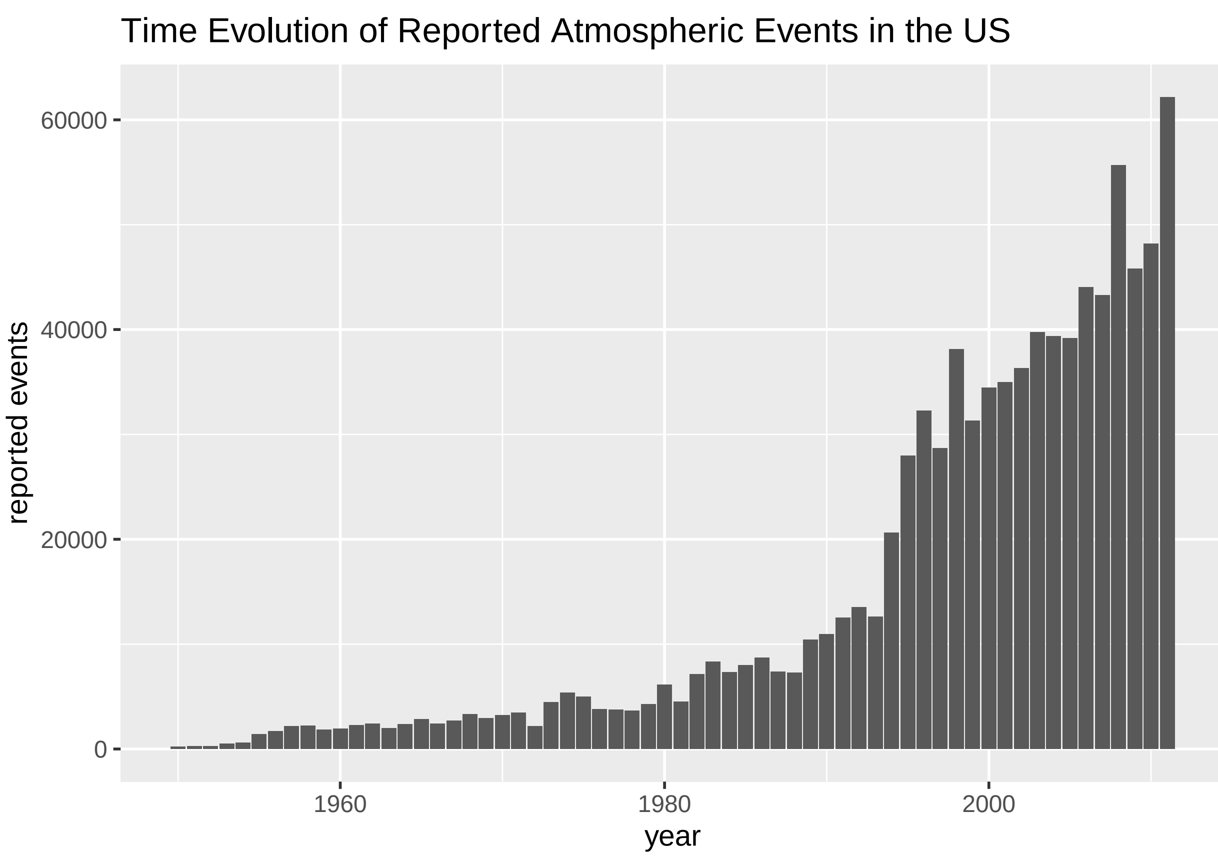

Figure 1: Reported events from 1950 to 2011

The ordered datasets can themselves deliver a pretty good summary of the situation in terms of fatalities:

```
data.fatalities[1:10, .(Type, Total.fatalities)]
```

```
##                Type Total.fatalities
##  1:         TORNADO             5633
##  2:  EXCESSIVE HEAT             1903
##  3:     FLASH FLOOD              978
##  4:            HEAT              937
##  5:       LIGHTNING              816
##  6:       TSTM WIND              504
##  7:           FLOOD              470
##  8:     RIP CURRENT              368
##  9:       HIGH WIND              248
## 10:       AVALANCHE              224
```

and injuries:

```
data.injuries[1:10, .(Type, Total.injuries)]
```

```
##                   Type Total.injuries
##  1:            TORNADO          91346
##  2:          TSTM WIND           6957
##  3:              FLOOD           6789
##  4:     EXCESSIVE HEAT           6525
##  5:          LIGHTNING           5230
##  6:               HEAT           2100
##  7:          ICE STORM           1975
##  8:        FLASH FLOOD           1777
##  9: THUNDERSTORM WIND           1488
## 10:               HAIL           1361
```

We can the plot the results:

```
# create the plots
library(gridExtra)
g1 <- ggplot(data = data.fatalities[1:10,], aes(x = Type, y = Perc.fatalities)) +
      geom_bar(stat = "identity") +
      xlab("") +
      ylab("Fraction of fatalities on total no.") +
      scale_x_discrete(limits = data.fatalities$Type[1:10]) +
      theme(axis.text.x = element_text(angle = 45, hjust = 1))
g2 <- ggplot(data = data.injuries[1:10,], aes(x = Type, y = Perc.injuries)) +
      geom_bar(stat = "identity") +
      xlab("") +
      ylab("Fraction of injuries on total no.") +
      scale_x_discrete(limits = data.injuries$Type[1:10]) +
      theme(axis.text.x = element_text(angle = 45, hjust = 1))
grid.arrange(g1, g2, ncol = 2)
```

It seems therefore that across the US, the largest impact on public health was due to tornado events (and wind-related reports) with a minor component correlated with excessive heat and floods.

Impact on Economy

The same kind of analysis can be performed on the economic consequences of severe weather conditions. In this case we analyse data on **property** and **crop** damage:

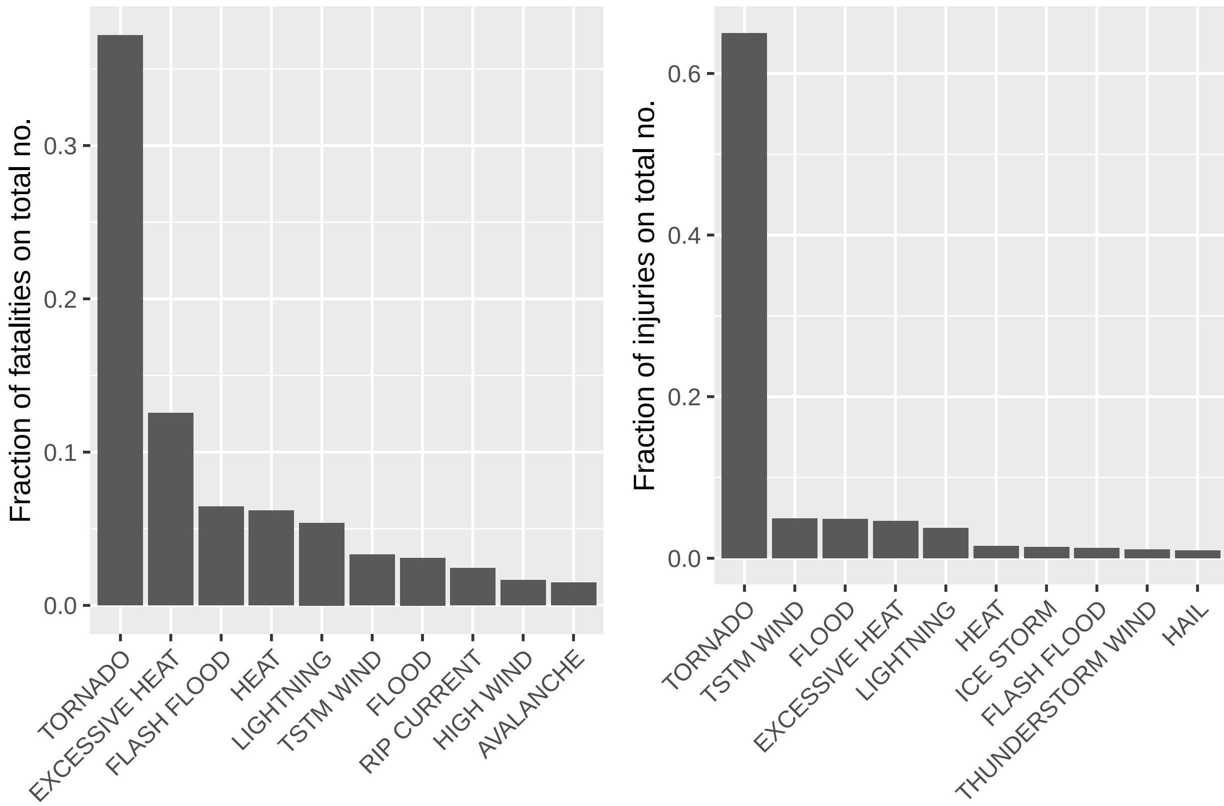

Figure 2: Casualties and injuries due to severe weather conditions from 1950 to 2011 in the US

```
data.type2 <- data.analysis[, .(Total.property.damage = sum(Property.damage),
                                Total.crop.damage     = sum(Crop.damage)
                               ),
                             by = Type
                            ]

# other than the total amount, we add the percentage of property and crop damage
data.type2[, Perc.property.damage := Total.property.damage / sum(Total.property.damage)]
data.type2[, Perc.crop.damage     := Total.crop.damage / sum(Total.crop.damage)]
```

As before, we rank the 10 most important causes of damage and plot the results:

```
# order by percentage
data.property <- data.type2[order(-data.type2$Total.property.damage),]
data.crop     <- data.type2[order(-data.type2$Total.crop.damage),]
```

which can already be a good metric of the analysis in terms of property damage (in billions of US$):

```
data.property[1:10, .(Type, Total.property.damage)]
```

```
##                  Type Total.property.damage
##  1:           TORNADO             3212258.2
##  2:       FLASH FLOOD             1420124.6
##  3:         TSTM WIND             1335965.6
##  4:             FLOOD              899938.5
##  5:  THUNDERSTORM WIND              876844.2
```

```
##  6:              HAIL              688693.4
##  7:         LIGHTNING              603351.8
##  8: THUNDERSTORM WINDS              446293.2
##  9:         HIGH WIND              324731.6
## 10:      WINTER STORM              132720.6
```

and crop damage (in billions of US$):

```
data.crop[1:10, .(Type, Total.crop.damage)]
```

```
##                  Type Total.crop.damage
##  1:              HAIL         579596.28
##  2:       FLASH FLOOD         179200.46
##  3:             FLOOD         168037.88
##  4:         TSTM WIND         109202.60
##  5:           TORNADO         100018.52
##  6:  THUNDERSTORM WIND          66791.45
##  7:           DROUGHT          33898.62
##  8: THUNDERSTORM WINDS          18684.93
##  9:         HIGH WIND          17283.21
## 10:        HEAVY RAIN          11122.80
```

Another interesting detail can be the fraction of the cost of weather factors with respect to the top element (i.e. normalised to the top cause of expenses) for property damage:

```
data.property[1:10, .(Type, Perc.cost = Total.property.damage / data.property$Total.property.damage[1])]
```

```
##                  Type  Perc.cost
##  1:           TORNADO 1.00000000
##  2:       FLASH FLOOD 0.44209541
##  3:         TSTM WIND 0.41589609
##  4:             FLOOD 0.28015758
##  5:  THUNDERSTORM WIND 0.27296815
##  6:              HAIL 0.21439540
##  7:         LIGHTNING 0.18782792
##  8: THUNDERSTORM WINDS 0.13893441
##  9:         HIGH WIND 0.10109136
## 10:      WINTER STORM 0.04131691
```

and crop damage:

```
data.crop[1:10, .(Type, Perc.cost = Total.crop.damage / data.crop$Total.crop.damage[1])]
```

```
##                  Type  Perc.cost
##  1:              HAIL 1.00000000
##  2:       FLASH FLOOD 0.30918152
##  3:             FLOOD 0.28992229
##  4:         TSTM WIND 0.18841149
##  5:           TORNADO 0.17256584
##  6:  THUNDERSTORM WIND 0.11523789
##  7:           DROUGHT 0.05848661
##  8: THUNDERSTORM WINDS 0.03223784
##  9:         HIGH WIND 0.02981939
## 10:        HEAVY RAIN 0.01919060
```

We finally plot the results:

```
# create the plots
g1 <- ggplot(data = data.property[1:10,], aes(x = Type, y = Total.property.damage)) +
      geom_bar(stat = "identity") +
      xlab("") +
      ylab("Property damage [billion of US$]") +
      scale_x_discrete(limits = data.property$Type[1:10]) +
      theme(axis.text.x = element_text(angle = 45, hjust = 1))
g2 <- ggplot(data = data.crop[1:10,], aes(x = Type, y = Total.crop.damage)) +
      geom_bar(stat = "identity") +
      xlab("") +
      ylab("Crop damage [billion of US$]") +
      scale_x_discrete(limits = data.crop$Type[1:10]) +
      theme(axis.text.x = element_text(angle = 45, hjust = 1))
grid.arrange(g1, g2, ncol = 2)
```

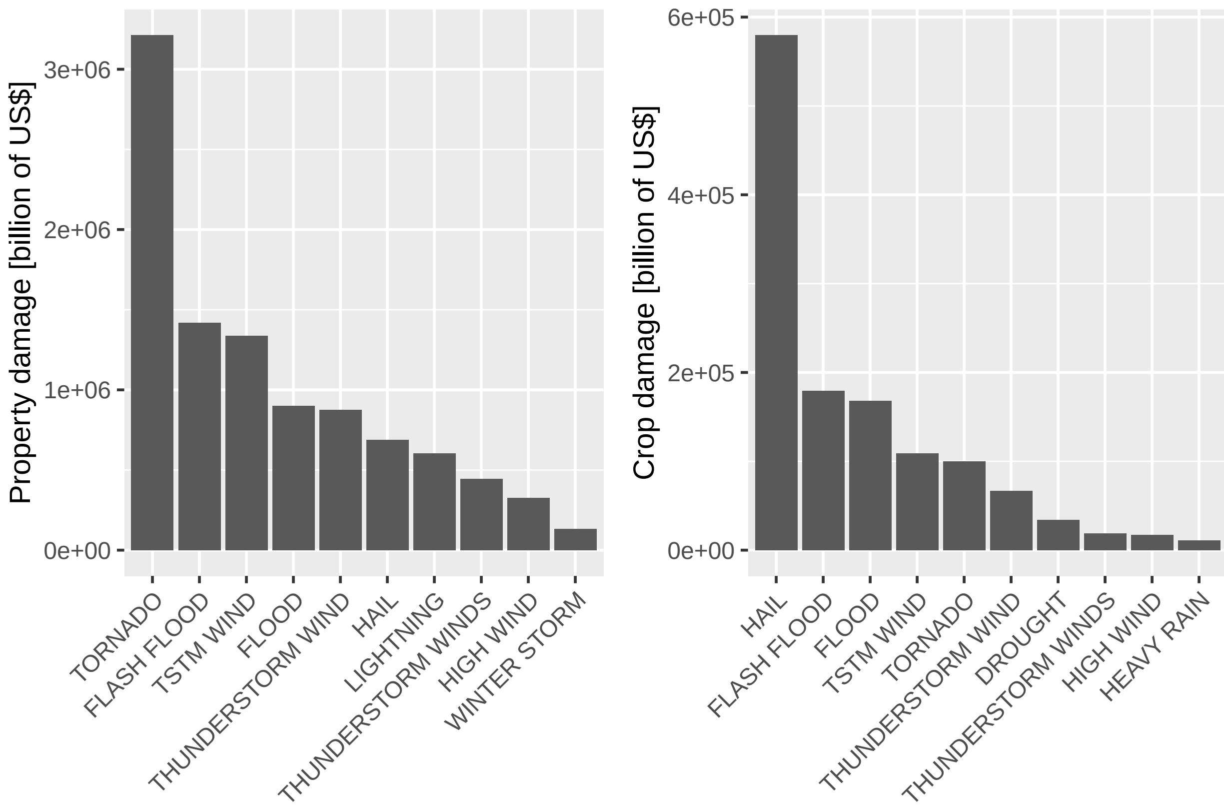

Figure 3: Property and crop damage due to severe weather conditions from 1950 to 2011 in the US

It seems therefore that tornado events are again responsible for a large part of the damage to properties, while the floods and hail have definitely a larger impact on agricultural-related damage.

Conclusions

The study is definitely not conclusive, but it may suggest that tornado and high speed winds play a central role in producing damage and public health issues, while most other factors have definitely more marginal parts. In terms of crop damage hail and heavy rain leading to floods seem to be the cause of most of the damage, leading the US government to spend more than threetimes the money for hail than for floods.