NOAA Storm Data Analysis

Yuling Tu

September 17, 2017

Severe Weather Events on Polulation Health and Economy

Synopsis

The U.S. National Oceanic and Atmospheric Administration's (NOAA) storm 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.

This report explores the NOAA Storm data and identify the impact of severe weather events on population health and economic consequences. This data set is collected from 1995 to November 2011.

loading and processing raw data

Download the data set

```
if(!file.exists("repdata_data_StormData.csv.bz2")) {
        temp <- tempfile()

download.file("https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormDa
ta.csv.bz2",temp)
        unlink(temp)
}
```

process the data

```
stormdata <- read.csv(file = "repdata_data_StormData.csv.bz2")
```

review the data structure and samples

```
str(stormdata)

## 'data.frame':    902297 obs. of  37 variables:
##  $ STATE__   : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ BGN_DATE  : Factor w/ 16335 levels "1/1/1966 0:00:00",..: 6523 6523
4242 11116 2224 2224 2260 383 3980 3980 ...
##  $ BGN_TIME  : Factor w/ 3608 levels "00:00:00 AM",..: 272 287 2705 1683
2584 3186 242 1683 3186 3186 ...
##  $ TIME_ZONE : Factor w/ 22 levels "ADT","AKS","AST",..: 7 7 7 7 7 7 7 7 7
7 ...
##  $ COUNTY    : num  97 3 57 89 43 77 9 123 125 57 ...
##  $ COUNTYNAME: Factor w/ 29601 levels "","5NM E OF MACKINAC BRIDGE TO
PRESQUE ISLE LT MI",..: 13513 1873 4598 10592 4372 10094 1973 23873 24418
4598 ...
##  $ STATE     : Factor w/ 72 levels "AK","AL","AM",..: 2 2 2 2 2 2 2 2 2 2
...
##  $ EVTYPE    : Factor w/ 985 levels "   HIGH SURF ADVISORY",..: 834 834
834 834 834 834 834 834 834 834 ...
##  $ BGN_RANGE : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ BGN_AZI   : Factor w/ 35 levels "","  N"," NW",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ BGN_LOCATI: Factor w/ 54429 levels "","- 1 N Albion",..: 1 1 1 1 1 1 1
1 1 1 ...
##  $ END_DATE  : Factor w/ 6663 levels "","1/1/1993 0:00:00",..: 1 1 1 1 1 1
1 1 1 1 ...
##  $ END_TIME  : Factor w/ 3647 levels ""," 0900CST",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ COUNTY_END: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ COUNTYENDN: logi  NA NA NA NA NA NA ...
##  $ END_RANGE : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ END_AZI   : Factor w/ 24 levels "","E","ENE","ESE",..: 1 1 1 1 1 1 1 1
1 1 ...
##  $ END_LOCATI: Factor w/ 34506 levels "","- .5 NNW",..: 1 1 1 1 1 1 1 1 1
1 ...
##  $ LENGTH    : num  14 2 0.1 0 0 1.5 1.5 0 3.3 2.3 ...
##  $ WIDTH     : num  100 150 123 100 150 177 33 33 100 100 ...
##  $ F         : int  3 2 2 2 2 2 2 1 3 3 ...
##  $ MAG       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ FATALITIES: num  0 0 0 0 0 0 0 0 1 0 ...
##  $ INJURIES  : num  15 0 2 2 2 6 1 0 14 0 ...
##  $ PROPDMG   : num  25 2.5 25 2.5 2.5 2.5 2.5 2.5 25 25 ...
##  $ PROPDMGEXP: Factor w/ 19 levels "","-","?","+",..: 17 17 17 17 17 17 17
17 17 17 ...
##  $ CROPDMG   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ CROPDMGEXP: Factor w/ 9 levels "","?","0","2",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ WFO       : Factor w/ 542 levels ""," CI","$AC",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ STATEOFFIC: Factor w/ 250 levels "","ALABAMA, Central",..: 1 1 1 1 1 1
1 1 1 1 ...
##  $ ZONENAMES : Factor w/ 25112 levels "","
"| __truncated__,..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ LATITUDE  : num  3040 3042 3340 3458 3412 ...
##  $ LONGITUDE : num  8812 8755 8742 8626 8642 ...
##  $ LATITUDE_E: num  3051 0 0 0 0 ...
##  $ LONGITUDE_: num  8806 0 0 0 0 ...
##  $ REMARKS   : Factor w/ 436781 levels "","-2 at Deer Park\n",..: 1 1 1 1
1 1 1 1 1 1 ...
##  $ REFNUM    : num  1 2 3 4 5 6 7 8 9 10 ...
```

The storm data set has 902297 observations and 37 variables. In this report, only the following variables are used to analyze the impact on public health and economy.

- EVTYPE -- type of event
- FATALITIES -- number of death
- INJURIES -- number of injuries
- PROPDMG,PROPDMGEXP,CROPDMG & CROPDMGEXP -- estimate damages in dollars

Create another data set for those variables only.

```
fd <- stormdata[, c("EVTYPE", "FATALITIES", "INJURIES", "PROPDMG",
"PROPDMGEXP", "CROPDMG", "CROPDMGEXP")]
str(fd)
```

```
## 'data.frame':    902297 obs. of  7 variables:
##  $ EVTYPE    : Factor w/ 985 levels "   HIGH SURF ADVISORY",..: 834 834
834 834 834 834 834 834 834 834 ...
##  $ FATALITIES: num  0 0 0 0 0 0 0 0 1 0 ...
##  $ INJURIES  : num  15 0 2 2 2 6 1 0 14 0 ...
##  $ PROPDMG   : num  25 2.5 25 2.5 2.5 2.5 2.5 2.5 25 25 ...
##  $ PROPDMGEXP: Factor w/ 19 levels "","-","?","+",..: 17 17 17 17 17 17 17
17 17 17 ...
##  $ CROPDMG   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ CROPDMGEXP: Factor w/ 9 levels "","?","0","2",..: 1 1 1 1 1 1 1 1 1 1
...
```

Cleaning Data

EVTYPE, FATALITIES, INJURIES, PROPDMG & CROPDMG variables all have valid data and no NULL data.
However, the unit of PROPDMGEXP and CROPDMGEXP has unexpected data which are not explained in the NOAA document, such as "?" (7 records in CROPDMGEXP, 8 records in PROPDMGEXP), "-" (1 record in PROPDMGEXP) and "+" (5 records in PROPDMGEXP).

This is the explanation in NOAA document regarding to DMGEXP. "Estimates should be rounded to three significant digits, followed by an alphabetical character signifying the magnitude of the number, i.e., 1.55B for $1,550,000,000. Alphabetical characters used to signify magnitude include "K" for thousands, "M" for millions, and "B" for billions. If additional precision is available, it may be provided in the narrative part of the entry."

For this report, those records with "?", "-", and "+" are set to NA and are removed in the calculation. Otherwise, the damage is calculated with this formula -- (DMG * 10^DMGEXP). For character DMGEXP, the following is the conversion.

- B or b is 10^9
- M or m is 10^6
- K or k is 10^3
- H or h is 10^2

```r
#create the array for DMGEXP
dmgexp <- c("", "+", "-", "?", 0:8, "h", "H", "k", "K", "m", "M", "b", "B")
# convert to digit for calculation later
# set "+", "-", "?" to NA
digit <- c(0, NA, NA, NA, 0:8, 2, 2, 3, 3, 6, 6, 9, 9)
# create lookup table
dmgexplookup <- data.frame(dmgexp, digit)

# add two new columns for PROPDMGEXP & PROPDMGEXP accordingly upon the lookup
table
fd$PROPDMGEXP2 <- dmgexplookup[match(fd$PROPDMGEXP,dmgexplookup$dmgexp),2]
fd$CROPDMGEXP2 <- dmgexplookup[match(fd$CROPDMGEXP,dmgexplookup$dmgexp),2]
```

Calculate all the impacted data by event type (EVTYPE)

```r
# load the library
library(plyr)

## Warning: package 'plyr' was built under R version 3.3.3

# remove NA for DMGEXP in "?", "-", "+"
result <- ddply(fd, .(EVTYPE), summarize,
                ftsum=sum(FATALITIES),
                injsum=sum(INJURIES),
                propdmgsum=sum(PROPDMG*10^PROPDMGEXP2, na.rm = TRUE),
                cropdmgsum=sum(CROPDMG*10^CROPDMGEXP2, na.rm = TRUE),
                dmgall=sum(propdmgsum,cropdmgsum))
```

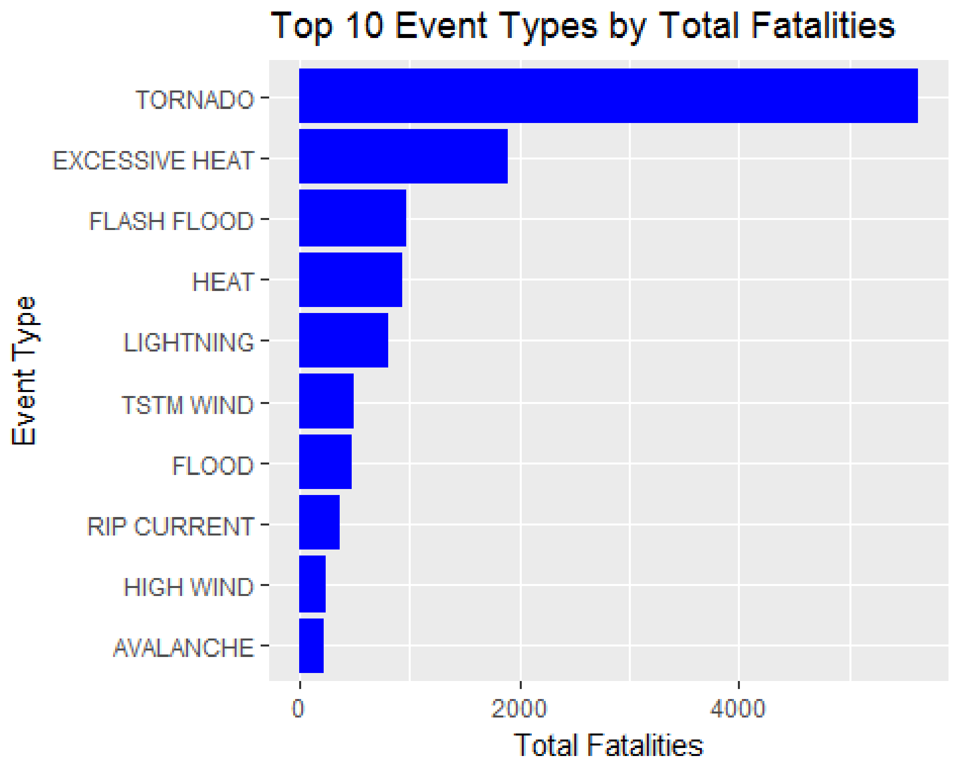
Analysis Results

Population Health Impact

For the population health impact, Fatalities and Injuries data table and charts are presented respectively below. Only top 10 events are displayed accordingly.

First, fatality trend is presented below.

```r
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.3.3

# get the top 10 event type by fatalities
ft <- result[order(result$ftsum, decreasing=TRUE),][1:10,]

# print top 10 Fatalities
print(ft[, c(1,2)], row.names = FALSE)

##          EVTYPE ftsum
##         TORNADO  5633
##  EXCESSIVE HEAT  1903
##     FLASH FLOOD   978
##             HEAT   937
##        LIGHTNING   816
##        TSTM WIND   504
##            FLOOD   470
##      RIP CURRENT   368
##        HIGH WIND   248
##        AVALANCHE   224

#plot the top 10
gf<-ggplot(ft, aes(y=ftsum, x=reorder(EVTYPE, ftsum)))+
  geom_bar(stat="identity", fill="blue")+
  coord_flip()+
  ggtitle("Top 10 Event Types by Total Fatalities")+
  xlab("Event Type")+
  ylab("Total Fatalities")

gf
```

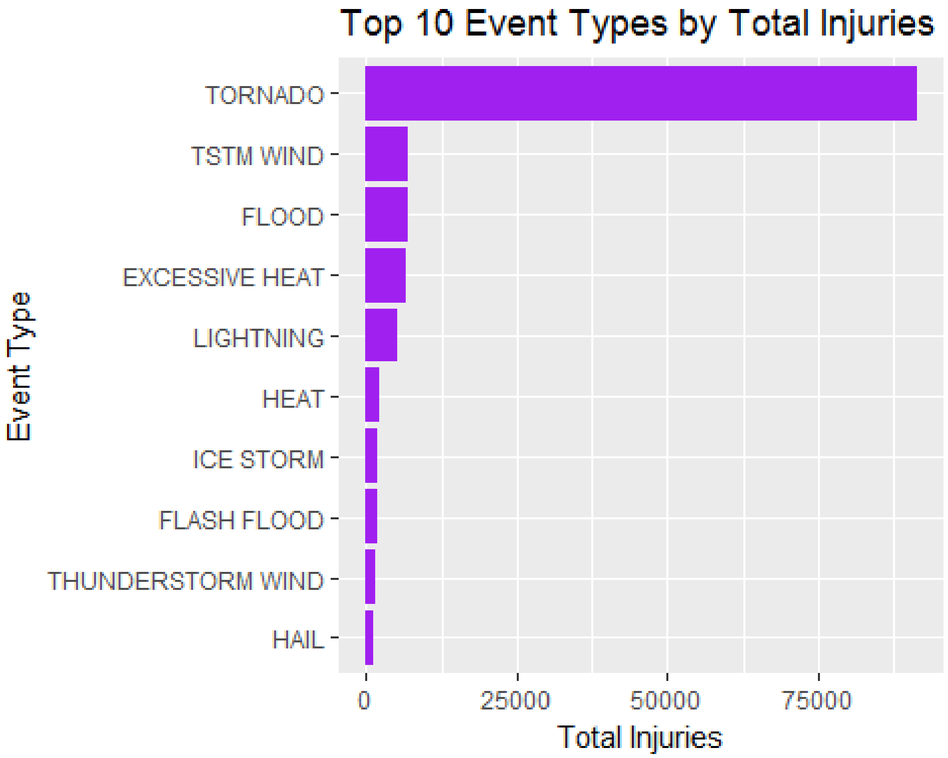


The event with the largest fatalities is "Tornado". The total of fatalities of Tornado is 5,633.

Second, injury trend is presented below.

```
# get the top 10 event type by injuries
inj <- result[order(result$injsum, decreasing=TRUE),][1:10,]

# print top 10 Injuries
print(ft[, c(1,3)], row.names = FALSE)
```

```
##           EVTYPE injsum
##          TORNADO  91346
##   EXCESSIVE HEAT   6525
##      FLASH FLOOD   1777
##             HEAT   2100
##        LIGHTNING   5230
##        TSTM WIND   6957
##            FLOOD   6789
##      RIP CURRENT    232
##        HIGH WIND   1137
##        AVALANCHE    170

#plot the top 10

gi<-ggplot(inj, aes(y=injsum, x=reorder(EVTYPE, injsum)))+
  geom_bar(stat="identity", fill="purple")+
  coord_flip()+
  ggtitle("Top 10 Event Types by Total Injuries")+
  xlab("Event Type")+
  ylab("Total Injuries")

gi
```



The event with the largest injuries is "Tornado", same as the fatalities. The total of injuries of Tornado is 91346.

From both charts of fatalities and injuries, it's very oblivious that Tornado is the most harmful event in US on population health. Top 2 events have big gap along with Tornado, especially on injuries.

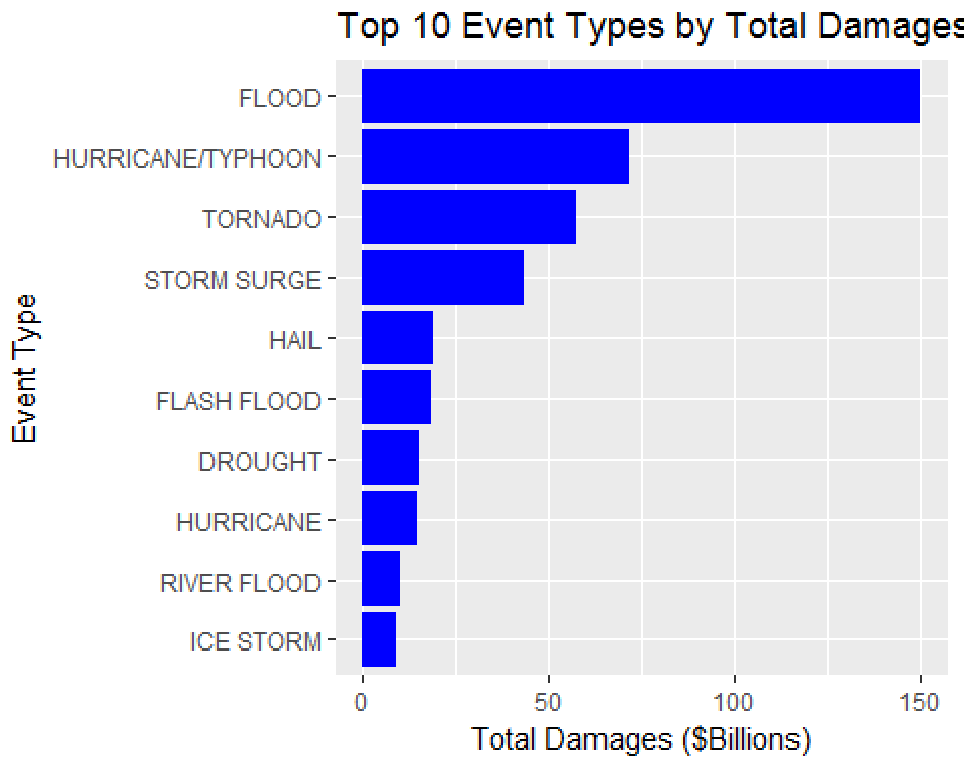
Economy Impact

For the economy impact, both property damage and crop damage are used as total damage in US dollars. Also, only top 10 event types are displayed.

```
# get the top 10 event type by total damages
totaldmg <- result[order(result$dmgall, decreasing=TRUE),][1:10,]

# print top 10 total damage
print(totaldmg[, c(1,6)], row.names = FALSE)
```

```
##             EVTYPE       dmgall
##              FLOOD 150319678257
##  HURRICANE/TYPHOON  71913712800
##            TORNADO  57362333887
##        STORM SURGE  43323541000
##               HAIL  18761221986
##        FLASH FLOOD  18243991079
##            DROUGHT  15018672000
##          HURRICANE  14610229010
##        RIVER FLOOD  10148404500
##          ICE STORM   8967041360
```

```
# from the top 10 data, all the damage are over billion. Use Billion for the
plot
#plot the top 10
gt<-ggplot(totaldmg, aes(y=dmgall/10^9, x=reorder(EVTYPE, dmgall)))+
  geom_bar(stat="identity", fill="blue")+
  coord_flip()+
  ggtitle("Top 10 Event Types by Total Damages")+
  xlab("Event Type")+
  ylab("Total Damages ($Billions)")
gt
```



From the chart, "Flood" caused the most damages and the estimation of damages is around 150 billion.