PA2-NOAA storm data analysis

ChenXiang

23 July 2015

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. In this report we analysed the NOAA Storm Database which tracks characteristics of major storms and weather events in the United States. We found that tornado, heat and flood are the 3 top harmful event type to people health. In specific, Since 1990 heat has caused over 3000 fatalies and tornado has caused over 25000 injuries. In terms of economic consquence, flood is the most harmful event to property damage which caused over ten billions dollar loss from 1990, followed by hurricane and tornado. Drought is the most destructive disaster to corps.

Motivation

The basic goal of this report is to explore the NOAA Storm Database and answer some basic questions about severe weather events.

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

Data Processing

Loading Data

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. From NOAA we obtain the storm database. 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. 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
- National Climatic Data Center Storm Events FAQ

We first read the data from the csv file included in the bzip archive. The data is a delimited file with fields delimited by the , character.

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

```
stormdata <- read.csv("repdata-data-StormData.csv.bz2", stringsAsFactors = FALSE, na.strings = c("NA", "
dim(stormdata)
```

```
## [1] 902297     37
```

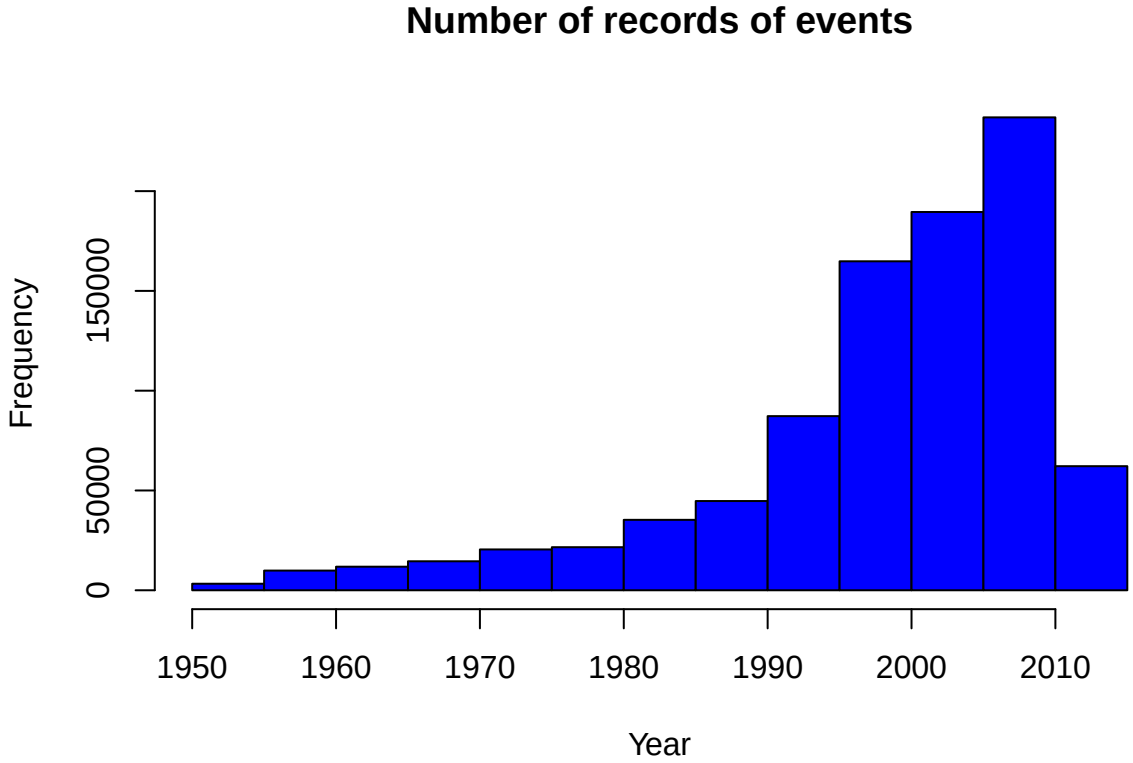
```
head(stormdata, 10)[, 1:10]
```

```
##    STATE__           BGN_DATE BGN_TIME TIME_ZONE COUNTY COUNTYNAME STATE
## 1        1  4/18/1950 0:00:00     0130       CST     97     MOBILE    AL
## 2        1  4/18/1950 0:00:00     0145       CST      3    BALDWIN    AL
## 3        1  2/20/1951 0:00:00     1600       CST     57    FAYETTE    AL
## 4        1   6/8/1951 0:00:00     0900       CST     89    MADISON    AL
## 5        1 11/15/1951 0:00:00     1500       CST     43    CULLMAN    AL
## 6        1 11/15/1951 0:00:00     2000       CST     77 LAUDERDALE    AL
## 7        1 11/16/1951 0:00:00     0100       CST      9     BLOUNT    AL
## 8        1  1/22/1952 0:00:00     0900       CST    123 TALLAPOOSA    AL
## 9        1  2/13/1952 0:00:00     2000       CST    125 TUSCALOOSA    AL
## 10       1  2/13/1952 0:00:00     2000       CST     57    FAYETTE    AL
##     EVTYPE BGN_RANGE BGN_AZI
## 1  TORNADO         0    <NA>
## 2  TORNADO         0    <NA>
## 3  TORNADO         0    <NA>
## 4  TORNADO         0    <NA>
## 5  TORNADO         0    <NA>
## 6  TORNADO         0    <NA>
## 7  TORNADO         0    <NA>
## 8  TORNADO         0    <NA>
## 9  TORNADO         0    <NA>
## 10 TORNADO         0    <NA>
```

Processing Data

Here we do some preprocessing for later analysis.

```
stormdata$BGN_DATE <- as.Date(stormdata$BGN_DATE, "%m/%d/%Y")
hist(as.numeric(format(stormdata$BGN_DATE, "%Y")), col = "blue", xlab = "Year", main = "Number of recor
```



From the above figure we know the event records before 1990 is much less than those after 1990. So we choose the data records after 1990. It will be enough for analysing the data set.

```
stormdata <- stormdata[stormdata$BGN_DATE >= as.Date("1/1/1990", "%m/%d/%Y"), ]
```

There are many alias for the `EVTYPE` variable. We need to rename them to the exact event type name defined by NOAA. I think it's a hard work. For now I can not figure out a good method but to use `grep` to deal with it.

```
event_types <- c("Astronomical Low Tide" ,"Avalanche", "Blizzard", "Coastal Flood", "Cold Chill", "Wind
type_flag <- logical(nrow(stormdata))
for (i in 1:length(event_types)) {
  matchindex <- grep(event_types[i], stormdata$EVTYPE, ignore.case = TRUE ,value = FALSE)
  stormdata$EVTYPE[matchindex] <- event_types[i]
  type_flag[matchindex] <- TRUE
}
```

There are many event types are not recognised, we just ignore them.

```
stormdata <- stormdata[type_flag, ]
stormdata$EVTYPE <- factor(stormdata$EVTYPE)
stormdata$PROPDMGEXP <- factor(stormdata$PROPDMGEXP)
stormdata$CROPDMGEXP <- factor(stormdata$CROPDMGEXP)
```

Results

which types of events are most harmful with respect to population health

1. Figure out the fatalities and injuries caused by each type of events.

```
pophealthdata <- aggregate(cbind(FATALITIES, INJURIES) ~ EVTYPE, data = stormdata, sum)
```

2. For simplicity just keep the top 10 event types to fatalities and injuries.

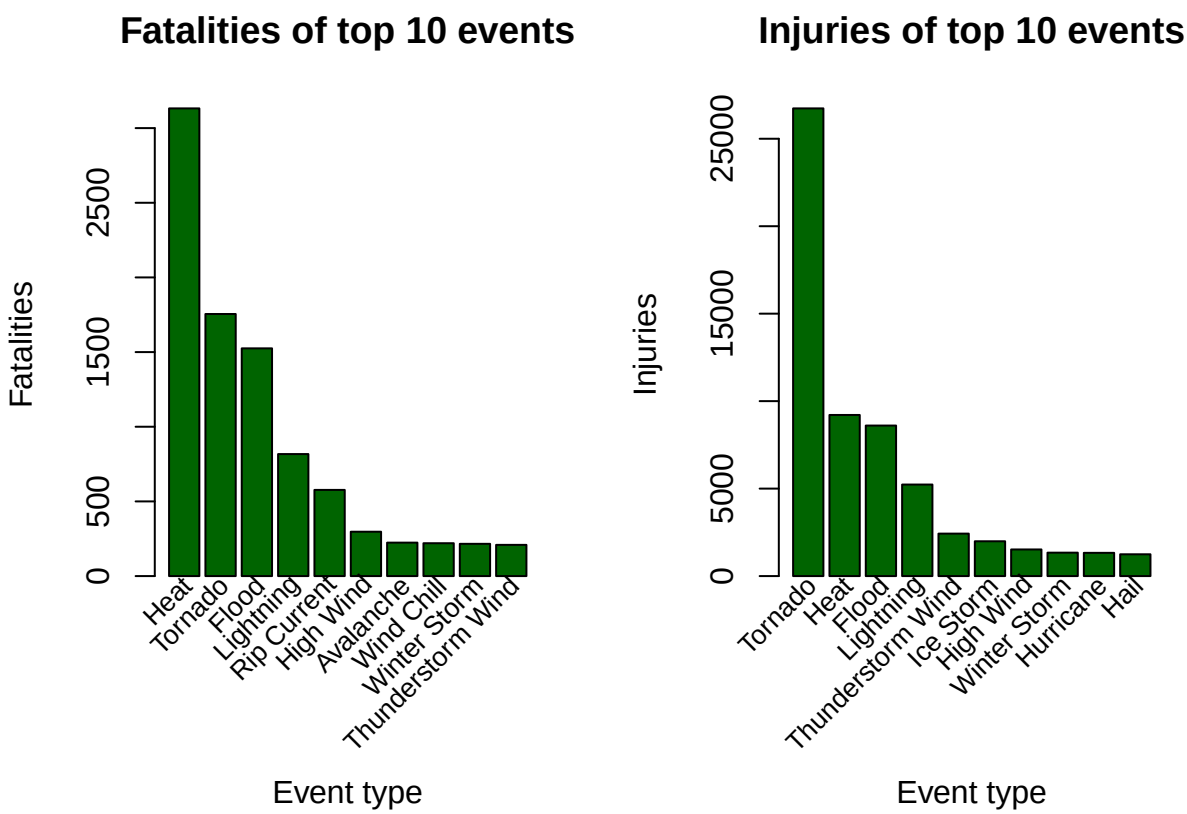
```
top10fatalities <- pophealthdata[order(pophealthdata$FATALITIES, decreasing = TRUE)[1:10], c(1, 2)]
top10injuries <- pophealthdata[order(pophealthdata$INJURIES, decreasing = TRUE)[1:10], c(1, 3)]
```

3. Make a figure to show the fatalities and injuries of top 10 different event types.

```
par(mfrow = c(1, 2))
par(mar = c(6, 4, 4, 2) + 0.1)

barplot(top10fatalities$FATALITIES, col = "darkgreen", ylab = "Fatalities", main = "Fatalities of top 1
text(seq(from = 0.8, by = 1.2, length.out = 10), par("usr")[3], srt = 45, adj = 1, cex = 0.85, labels =
mtext(1, text = "Event type", line = 5)

barplot(top10injuries$INJURIES, col = "darkgreen", ylab = "Injuries", main = "Injuries of top 10 events
text(seq(from = 0.8, by = 1.2, length.out = 10), par("usr")[3], srt = 45, adj = 1, cex = 0.85, labels =
mtext(1, text = "Event type", line = 5)
```



As illustrated by the figure, from the aspect of fatalities, heat is the most harmful event type and killed over 3000 person from 1990. But tornado leads to most injuries ,over 20000 from 1990, which considerably exceeds the follows. Take both into consideration, Tornado, Heat and Flood(include all kinds of flood) are the top 3 harmful event types.

Which types of events have the greatest economic consequences

1. Calculate the economic loss of different event types.

```
stormdata$PROPDMGEXP <- toupper(stormdata$PROPDMGEXP)
stormdata$CROPDMGEXP <- toupper(stormdata$CROPDMGEXP)
dollarunits <- c("H", "K", "M", "B")
propdmgdata <- stormdata[stormdata$PROPDMGEXP %in% dollarunits , ]
propdmgdata <- aggregate(PROPDMG ~ EVTYPE + PROPDMGEXP, data = propdmgdata, sum)
for(i in 1:nrow(propdmgdata)){
  propdmgdata$PROPDMG[i] <- if (propdmgdata$PROPDMGEXP[i] == "H") {
    propdmgdata$PROPDMG[i] * 100
  }else if(propdmgdata$PROPDMGEXP[i] == "K"){
    propdmgdata$PROPDMG[i] * 1000
  }else if(propdmgdata$PROPDMGEXP[i] == "M"){
    propdmgdata$PROPDMG[i] * 1000000
  }else if(propdmgdata$PROPDMGEXP[i] == "B"){
    propdmgdata$PROPDMG[i] * 1000000000
  }
}

cropdmgdata <- stormdata[stormdata$CROPDMGEXP %in% dollarunits , ]
cropdmgdata <- aggregate(CROPDMG ~ EVTYPE + CROPDMGEXP, data = cropdmgdata, sum)
for(i in 1:nrow(cropdmgdata)){
  cropdmgdata$CROPDMG[i] <- if (cropdmgdata$CROPDMGEXP[i] == "H") {
    cropdmgdata$CROPDMG[i] * 100
  }else if(cropdmgdata$CROPDMGEXP[i] == "K"){
    cropdmgdata$CROPDMG[i] * 1000
  }else if(cropdmgdata$CROPDMGEXP[i] == "M"){
    cropdmgdata$CROPDMG[i] * 1000000
  }else if(cropdmgdata$CROPDMGEXP[i] == "B"){
    cropdmgdata$CROPDMG[i] * 1000000000
  }
}
```

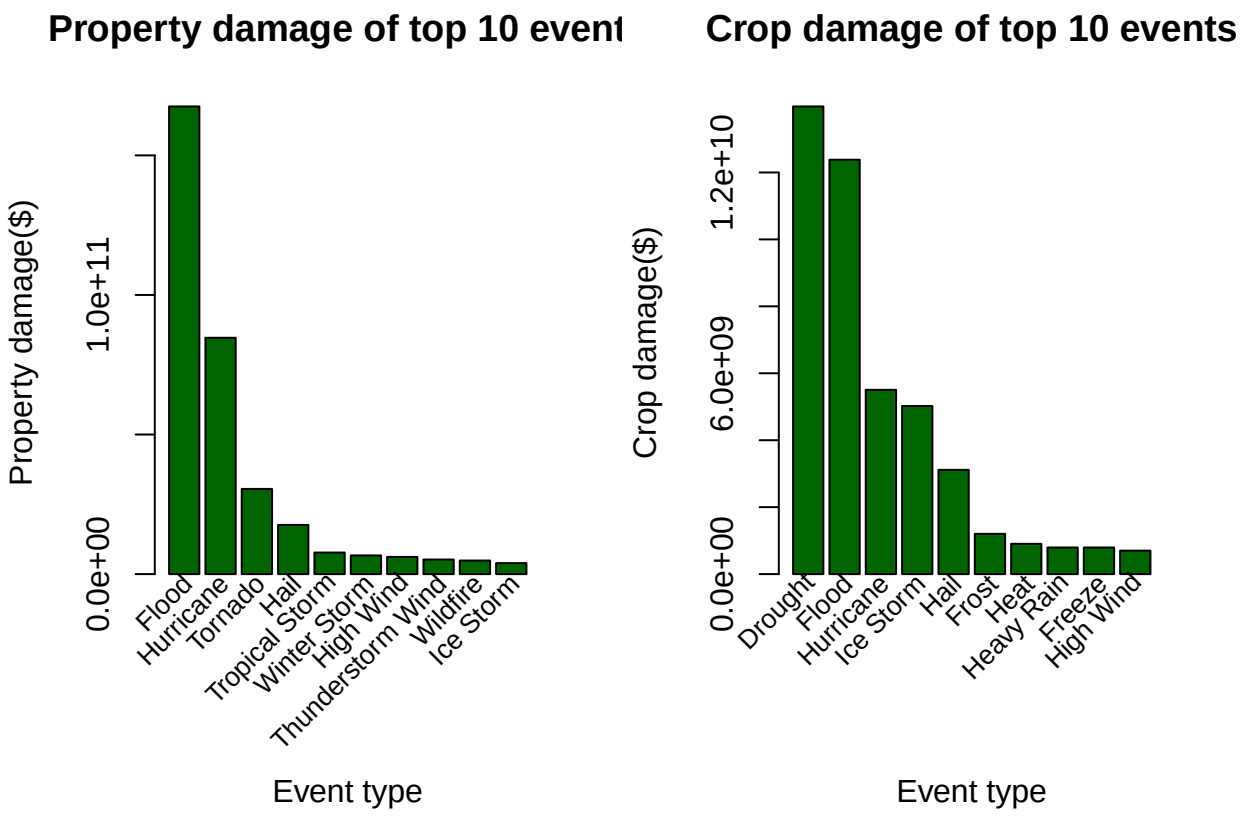
2. For simplicity just keep the top 10 event types to property and crop damage.

```
propdmgdata <- aggregate(PROPDMG ~ EVTYPE, data = propdmgdata, sum)
top10propdmg <- propdmgdata[order(propdmgdata$PROPDMG, decreasing = TRUE)[1:10], ]
cropdmgdata <- aggregate(CROPDMG ~ EVTYPE, data = cropdmgdata, sum)
top10cropdmg <- cropdmgdata[order(cropdmgdata$CROPDMG, decreasing = TRUE)[1:10], ]
```

3. Make a figure to show the fatalities and injuries of top 10 different event types.

```
par(mfrow = c(1, 2))
par(mar = c(6, 4, 4, 2) + 0.1)
barplot(top10propdmg$PROPDMG, col = "darkgreen", ylab = "Property damage($)", main = "Property damage o
text(seq(from = 0.8, by = 1.2, length.out = 10), par("usr")[3], srt = 45, adj = 1, cex = 0.85, labels =
mtext(1, text = "Event type", line = 5)

barplot(top10cropdmg$CROPDMG, col = "darkgreen", ylab = "Crop damage($)", main = "Crop damage of top 10
text(seq(from = 0.8, by = 1.2, length.out = 10), par("usr")[3], srt = 45, adj = 1, cex = 0.85, labels =
mtext(1, text = "Event type", line = 5)
```



As illustrated by the figure, flood, hurricane and tornado are the top 3 harmful event to property damage. But in terms of crop damage, drought is no doubt the most destructive, which can lead over 10 billions economic loss. Flood and hurricane are also more destructive to crops when compared to other disasters.