Which Storms Are The Most Harmful?

Steffen Hartleib

May 18, 2015

Summary

To identify the storms that cause the most health and economic damage I analzyed a data set from U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database. The data tracks major storm events in the United States, including when and where they occur, as well as estimates of any fatalities, injuries, and property damage.
I used the fatalities as a measure of health impact and the sum of crop and property damage as as the measure of economic impact.

15 Storm events caused 80% of the Fatalities while 7 Storm events caused 80% of the economic damge. Tornados and Flash Floods are in the top three of both lists.

Processing

I reduced the data set to the five relevant columns: Event Type, Beginning Date, Fatalities, Property Damage, and Crop damage. I also add a "Year" column to make it easier to track events and their impact over time.

```r
require(dplyr)
require(reshape2)
require(ggplot2)
options(scipen=999)
storm <- read.csv(bzfile("repdata-data-StormData.csv.bz2"))
storm <- select(storm,EVTYPE,BGN_DATE,FATALITIES,CROPDMG,PROPDMG)
storm$BGN_DATE <- as.Date(storm$BGN_DATE, format = "%m/%d/%Y %H:%M:%S")
storm <- mutate(storm, Year = as.POSIXlt(storm$BGN_DATE)$year  + 1900)
```

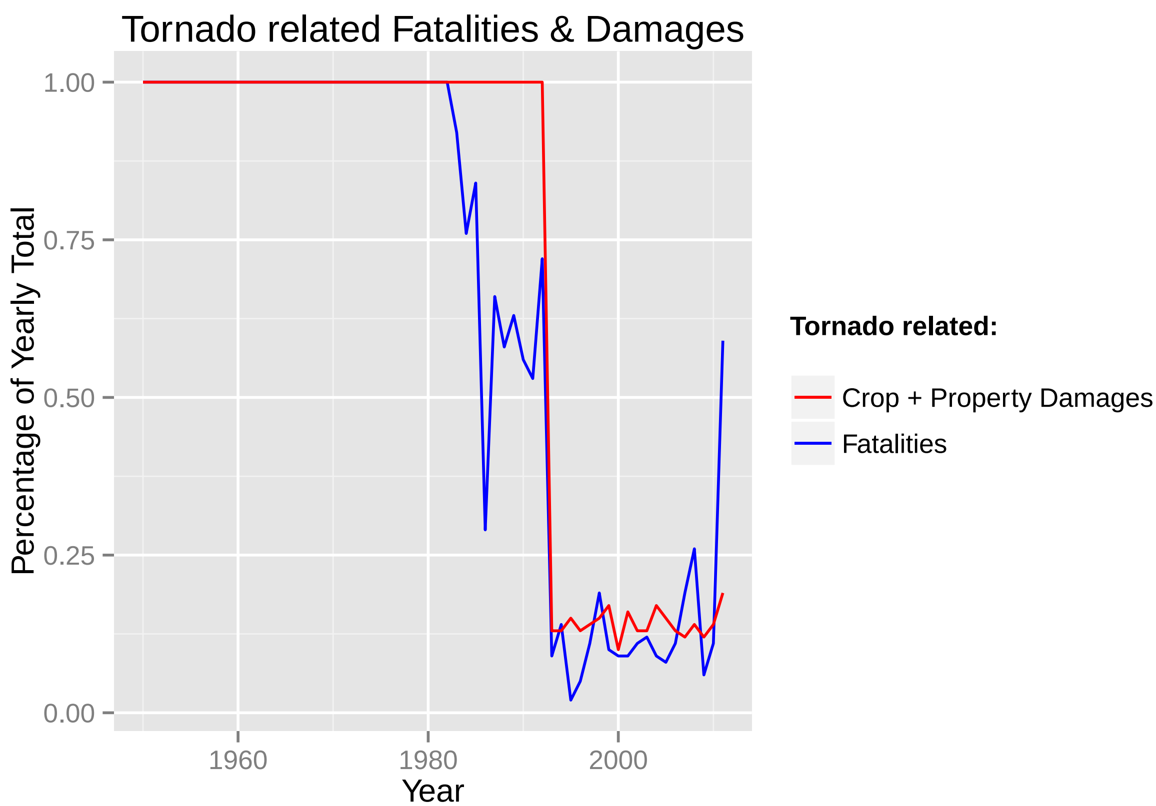
The data set spans from 1950 to 2011. But between 1950 and 1983 the data measures only tornado related fatalities. Between 1983 and 1993 the Tornado related fatalities drop as as a percentage of total, as data from other storm types are recorded. Starting 1993 they're fairly constant for the next 20 years. Damages behave similiarly. Until 1993 the data only shows Tornado related damages. Since it's very unlikely that Tornados became about 80% less amaging during the 80s I assumed that this early Tornado bias exists because data from other storms wasn't collected before 1983 and 1993 respectively. To allow for a meaningful comparison of Storm Types I excluded data from before 1993.

Figure 1

```r
# Tornado related fataltities and damages per year
TorFatDamYear <- filter(storm, storm$EVTYPE == "TORNADO")
TorFatDamYear <- group_by(TorFatDamYear, Year)
TorFatDamYear <- summarize(TorFatDamYear,
                          TorFat = sum(FATALITIES),
                          TorDam = sum(CROPDMG + PROPDMG))
# Total fatalites and damages per year
TotFatDamYear <- group_by(storm, Year)
```

```
TotFatDamYear <- summarize(TotFatDamYear,
                          TotFat = sum(FATALITIES),
                          TotDam = sum(CROPDMG + PROPDMG))
# Damagas and fatalities as percentage of yearly totals
TorVsTotYear <- data.frame(TotFatDamYear, TorFatDamYear[,-1])
TorVsTotYear <- mutate(TorVsTotYear, TornFatVsTot = round(TorFat/TotFat,2),
                          TornDamVsTot = round(TorDam/TotDam,2))
TorVsTotYear <- select(TorVsTotYear, Year, TornFatVsTot, TornDamVsTot)

ggplot(TorVsTotYear, aes(Year)) +
                        geom_line(aes(y = TornFatVsTot, colour = "TornFatVsTot")) +
                        geom_line(aes(y = TornDamVsTot, color = "TornDamVsTot")) +
                        labs(title = "Tornado related Fatalities & Damages", x = "Year",
                        y = "Percentage of Yearly Total" ) + scale_color_manual("Tornado related:\n", label
```



Results

Health Impact

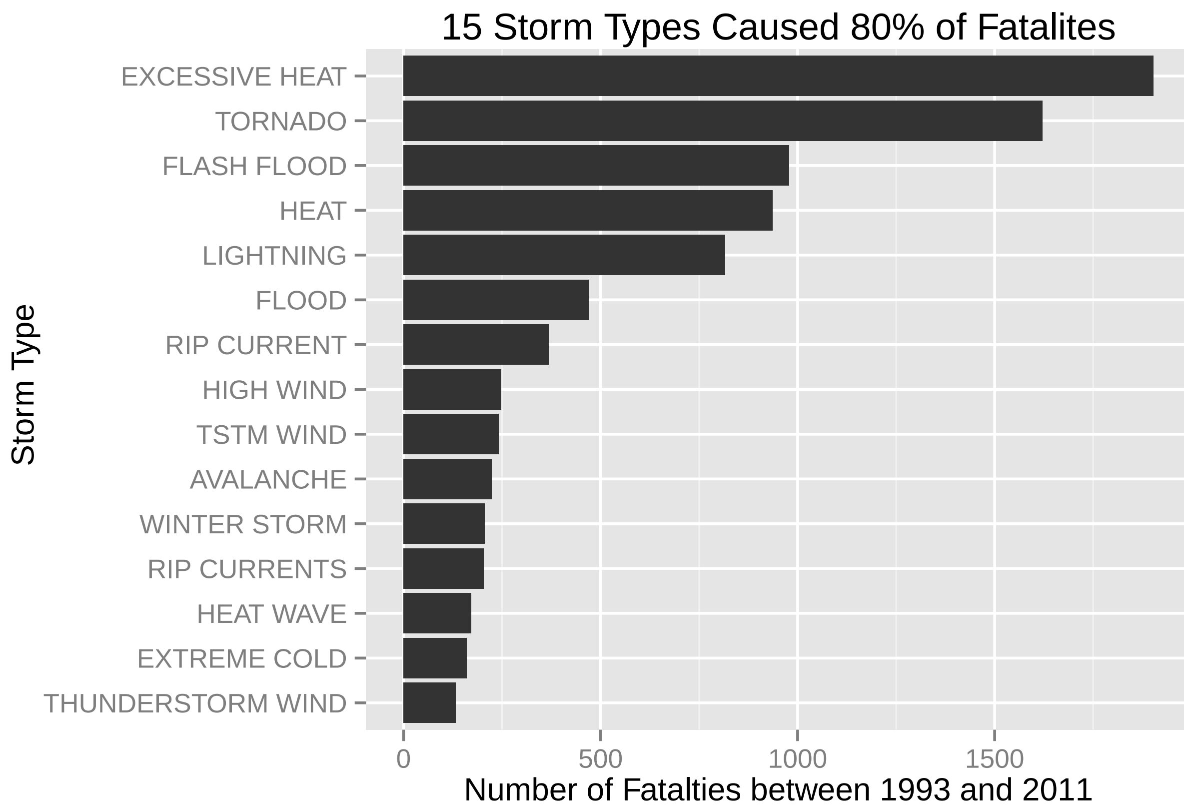
I used the total number of fatalities per Event Type as the measure of health impact. To answer the question "which storms are the most harmful", I identified the set of most harmful Event Types that caused 80% of the total damage. The table below shows Event Types ranked by total Fatalities, along with a column showing for each Event Type the percentage of Total Fatalities, and the cumulative percentage of Total Fatalities. It turns out that 80% of the fatalties are caused by just 15 Event Types. The results are shown the in the following table.

```
storm93 <- filter(storm, Year > 1992)
FatTot <- sum(storm93$FATALITIES)
HealthImpact <- select(storm93, EVTYPE,FATALITIES)
HealthImpact <- group_by(HealthImpact, EVTYPE)
HealthImpact <- summarise(HealthImpact, TotFat = sum(FATALITIES),FatPercOfTotFat = round(sum(FATALITIES)
HealthImpact <- arrange(HealthImpact, - TotFat)
HealthImpact <- mutate(HealthImpact, CumeFatPercTotFat = cumsum(FatPercOfTotFat))
HealthImpact<- filter(HealthImpact, CumeFatPercTotFat <= .8)
HealthImpact
```

```
## Source: local data frame [15 x 4]
##
##               EVTYPE TotFat FatPercOfTotFat CumeFatPercTotFat
## 1     EXCESSIVE HEAT   1903            0.18              0.18
## 2            TORNADO   1621            0.15              0.33
## 3        FLASH FLOOD    978            0.09              0.42
## 4               HEAT    937            0.09              0.51
## 5          LIGHTNING    816            0.08              0.59
## 6              FLOOD    470            0.04              0.63
## 7        RIP CURRENT    368            0.03              0.66
## 8          HIGH WIND    248            0.02              0.68
## 9          TSTM WIND    241            0.02              0.70
## 10         AVALANCHE    224            0.02              0.72
## 11      WINTER STORM    206            0.02              0.74
## 12      RIP CURRENTS    204            0.02              0.76
## 13         HEAT WAVE    172            0.02              0.78
## 14      EXTREME COLD    160            0.01              0.79
## 15 THUNDERSTORM WIND    133            0.01              0.80
```

Figure 2

```
HealthImpact$EVTYPE <- reorder(HealthImpact$EVTYPE, HealthImpact$TotFat)
ggplot(HealthImpact, aes(x = EVTYPE, y = TotFat)) + geom_bar(stat = "Identity") + coord_flip() + labs(ti
```



Ecomonic Impact

I used the sum of crop and property damage as the measure of economic impact. I applied the same 80% logic as above. It turns out that Economic impact is even more concentrated than the Health impact. Only 7 Storm Event Types are causing 81% of the damage. The results are shown in the following table.

```
DamTot <- sum(storm93$CROPDMG + storm93$PROPDMG)
EconImpact <- select(storm93, EVTYPE,CROPDMG,PROPDMG)
EconImpact <- group_by(EconImpact, EVTYPE)
EconImpact <- summarise(EconImpact,
                        TotDam = round(sum(CROPDMG +PROPDMG),0),
                        DamPercOfTotDam = round(sum(CROPDMG +PROPDMG)/DamTot,2))
EconImpact <- arrange(EconImpact, - TotDam)
EconImpact <- mutate(EconImpact, CumeDamPercTotDam = cumsum(DamPercOfTotDam))
EconImpact <- filter(EconImpact, CumeDamPercTotDam <= .81)
EconImpact
```

```
## Source: local data frame [7 x 4]
##
##              EVTYPE  TotDam DamPercOfTotDam CumeDamPercTotDam
## 1       FLASH FLOOD 1599325            0.15              0.15
## 2           TORNADO 1487776            0.14              0.29
## 3         TSTM WIND 1445168            0.14              0.43
## 4              HAIL 1268290            0.12              0.55
## 5             FLOOD 1067976            0.10              0.65
## 6 THUNDERSTORM WIND  943636            0.09              0.74
## 7         LIGHTNING  606932            0.06              0.80
```

Figure 3

```
EconImpact$EVTYPE <- reorder(EconImpact$EVTYPE, EconImpact$TotDam)
ggplot(EconImpact, aes(x = EVTYPE, y = TotDam)) + geom_bar(stat = "Identity") + coord_flip() + labs(titl
```