Analysis of the NOAA Storm Database and Severe Weather events

Chris Shaw

14 March 2016

Synopsis

Introduction

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.

The government have requested this report to help prepare for severe weather events, including prioritisation of resources for different types of events. The following questions are addressed:

1. Across the United States, which types of events are most harmful with respect to population health?
2. Across the United States, which types of events have the greatest economic consequences?

The analysis is based upon 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.

Data Processing

A number of steps are required to prepare the data before analysis which are described below. Some decisions have been made during the processing which impact the results, so a justification is provided.

Acquiring the data

The raw data was provided as an internet url. The dataset is rather large, so is only downloaded if it doesn't already exist:

```
rawdata_url <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
rawdata_file <- "storm_data.csv.bz2"
rawdata_csv <- "storm_data.csv"

if(!file.exists(rawdata_file)) {
        download.file(url = rawdata_url, destfile = rawdata_file)
        bunzip2(filename = rawdata_file, destname = rawdata_csv)
}
storm_data <- read.csv(rawdata_csv)
```

The raw data contains 902,297 weather observations and 37 variables for each observation.

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 are more complete.

Cleaning and transforming the data

The first task is to create a smaller dataset with only those items relevant to the analysis. A new variable called **year** is created to assist with group of events by calender year later on.

```
#library(plyr)
events<-storm_data %>% select(EVTYPE, BGN_DATE, FATALITIES, INJURIES, PROPDMG,
                              PROPDMGEXP, CROPDMG, CROPDMGEXP, STATE)
events$BGN_DATE <- as.Date(events$BGN_DATE, "%m/%d/%Y %H:%M:%S")
events$year <- as.numeric(format(events$BGN_DATE, "%Y"))
```

Two transformations have been applied to the data. The first is combine the **DMG** and **DMGEXP** variables for crop and property damage. In the raw data, the first contains a number and the second contains the units of measurement in US dollars (k=thousand, m=million and b=billion). Some observations have other indicators which could not be reliably interpreted, so these are set to zero for this analysis.

The correct scaling factor is applied to these and multiplied to create a new **dmg_k** variable to give a consistent economic damage value in thousand dollar units.

```
# Convert crop damage into consistent units and multiply
events$CROPDMGEXP<-mapvalues(events$CROPDMGEXP,
                             from = c("",  "?", "0", "2", "B", "k", "K", "m", "M"),
                             to = c(0,0,0,0,1000000, 1, 1, 1000, 1000))
events$cropdmg_k<-events$CROPDMG * as.numeric(as.character(events$CROPDMGEXP))

# Convert property damage into consistent units and multiply
events$PROPDMGEXP<-mapvalues(events$PROPDMGEXP,
                             from = c("M", "m", "K", "B", "", "-", "?", "+", "0", "1", "2",
                                      "3", "4", "5", "6", "7", "8", "h", "H"),
                             to = c(1000, 1000, 1, 1000000, 0,0,0,0,0,0,0,
                                    0,0,0,0,0,0,0,0))
events$propdmg_k<-events$PROPDMG * as.numeric(as.character(events$PROPDMGEXP))
```

The second transformation is to relabel the names of some events. During exploratory analysis it became clear that some types of similar events are labelled differently due to spelling mistakes and multiple categorisations.

For the analysis in this report,a certain number of categories have been relabelled as shown below. There are over 900 different event types, and only a few key ones have been selected that are material in this report. For example, all the different types of flooding are categorised as Flood.

```
events$EVTYPE<-revalue(events$EVTYPE, c("TSTM WIND"="THUNDERSTORM","FLASH FLOOD"="FLOOD",
                                        "FLASH FLOOD/FLOOD"="FLOOD", "FLASH FLOODING"="FLOOD",
                                        "HIGH WINDS"="HIGH WIND", "RIVER FLOOD"="FLOOD",
                                        "THUNDERSTORM WIND"="THUNDERSTORM", "THUNDERSTORM WINDS"="THUNDE
                                        "TORNADOES, TSTM WIND, HAIL"="TORNADO", "EXTREME HEAT"="HEAT","W
                                        "DAMAGING FREEZE"="FREEZE", "HEAVY RAIN/SEVERE WEATHER"="HEAVY R
                                        "HURRICANE OPAL"="HURRICANE", "RIP CURRENTS"="RIP CURRENT", "EXT
                                        "HURRICANE/TYPHOON"="HURRICANE", "WILD FIRES"="WILDFIRE",
                                        "STORM SURGE/TIDE"="STORM SURGE", "HURRICANE ERIN"="HURRICANE",
                                        "FLOOD/FLASH FLOOD"="FLOOD",
                                        "HEAT WAVE"="HEAT", "River Flooding"="FLOOD",
                                        "EXCESSIVE HEAT"="HEAT", "FROST/FREEZE"="FREEZE",
                                        "COLD/WIND CHILL"="COLD", "EXTREME COLD"="COLD",
                                        "HEAVY SURF/HIGH SURF"="HIGH SURF"))
#detach("package:plyr", unload=TRUE)
```

Creating an analytic dataset

After cleaning and transforming the data,the last step is to summarise annually to be able to analyse the full public health and econmic effects.

The total injuries, fatalities, property and crop damage is calculated per year across all recorded event types, and then the percentage contribution from each event type can be calculated for that year

summaries per year. The * Create a summarised set of data by year and EVTYPE

```
annual_events <- events %>% group_by(year) %>%
                            mutate(total_annual_events = n(),
                                   total_annual_injuries = sum(INJURIES),
                                   total_annual_fatalities = sum(FATALITIES),
                                   total_annual_property_damage = sum(propdmg_k),
                                   total_annual_crop_damage = sum(cropdmg_k)) %>%
                            group_by(year, EVTYPE, total_annual_events,
                                                   total_annual_injuries,
                                                   total_annual_fatalities,
                                                   total_annual_property_damage,
                                                   total_annual_crop_damage)  %>%
                            summarise(number_of_events = n(),
                                   event_injuries = sum(INJURIES),
                                   event_fatalities = sum(FATALITIES),
                                   event_property_damage = sum(propdmg_k),
                                   event_crop_damage = sum(cropdmg_k)) %>%
                             mutate(event_injury_percent = ifelse(total_annual_injuries==0,0,event_inju
                                    event_fatality_percent = ifelse(total_annual_fatalities==0,0,event_
                                    event_property_damage_percent = ifelse(total_annual_property_damage=
                                    event_crop_damage_percent = ifelse(total_annual_crop_damage==0,0,ev

# Helper functions to summarise data frames which have a text column followed by
# numerical ones
# crunch keeps specified rows in a data.frame and rolls up the others into a single
# row with the label text.
crunch <- function(df,rowstokeep, rolleduptext,rowstosum=-(rowstokeep)) {
        totals=list()
        for (i in 2:ncol(df)) totals[i-1]<- sum(df[rowstosum,i])
        rbind(df[rowstokeep,],setNames(data.frame(rolleduptext, totals), names(df)))
}
# totalcrunch calls crunch twice - first to crunch the data frame and again to add
# a grand total row
totalcrunch <- function(df, rowstokeep, rolleduptext){
        crunch(crunch(df, rowstokeep, rolleduptext), 1:(length(rowstokeep)+1),
                          "GRAND TOTAL", 1:(length(rowstokeep)+1))
}

population_impact <- annual_events %>% group_by(EVTYPE) %>%
                summarise(events=sum(number_of_events),fatalities=sum(event_fatalities),
                          injuries=sum(event_injuries)) %>% arrange(desc(fatalities))

economic_impact <- annual_events %>% group_by(EVTYPE) %>%
                summarise(events=sum(number_of_events),property_damage=sum(event_property_damage),
                          crop_damage=sum(event_crop_damage)) %>% arrange(desc(property_damage))
```

```
top_ten_impacts <- cbind(totalcrunch(population_impact, 1:10, "(OTHERS)"),
                         totalcrunch(economic_impact , 1:10, "(OTHERS)"))
kable(top_ten_impacts, digits=0, format.args=list(big.mark = ','),
      caption = "Top 10 Public Health and economic impacts since 1950",
      col.names = c("Event", "Observations", "Fatalities", "Injuries",
                    "Event", "Observations", "Property Damage ($'000)", "Crop Damage ($'000)"))
```

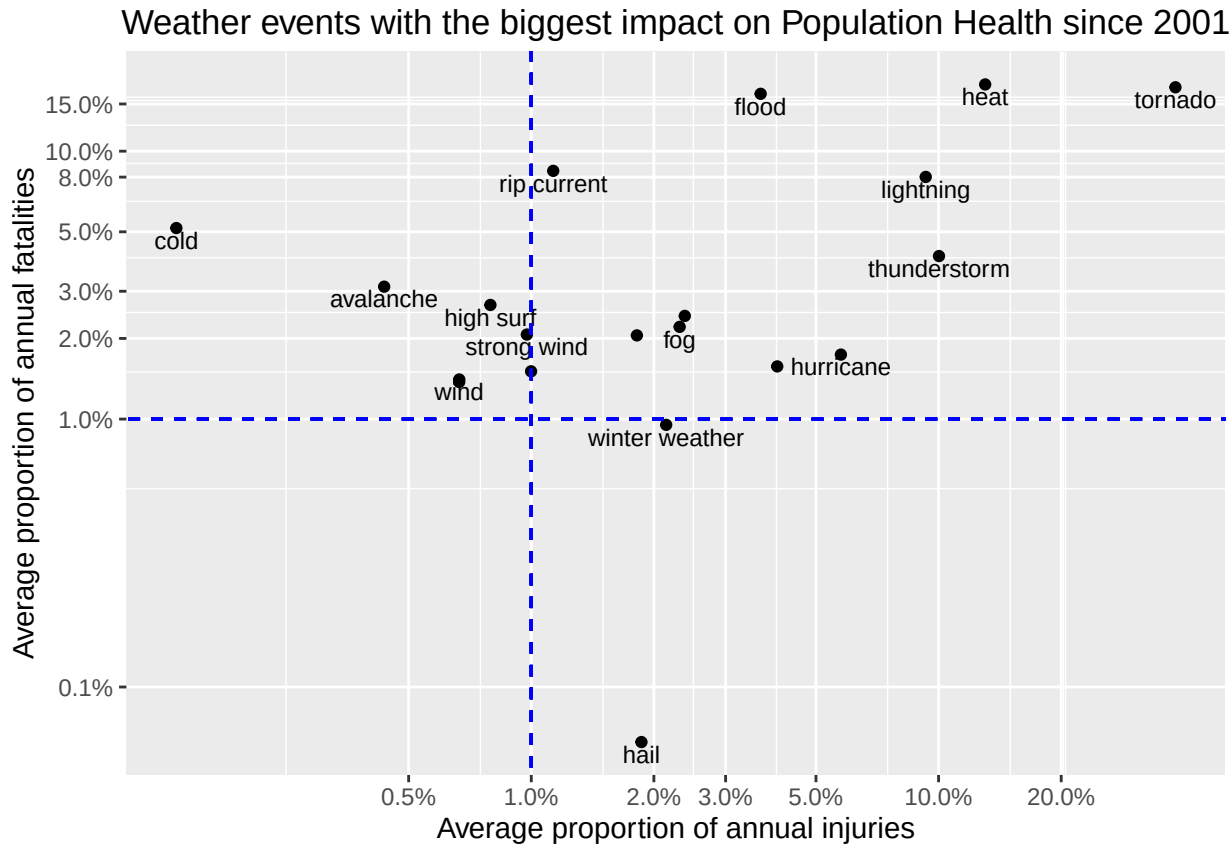
Table 1: Top 10 Public Health and economic impacts since 1950

Event	Observations	Fatalities	Injuries	Event	Observations	Property Damage ($'00
TORNADO	60,653	5,658	91,346	FLOOD	81,109	166,777,8
HEAT	2,541	3,108	9,089	HURRICANE	278	84,605,1
FLOOD	81,109	1,500	8,592	TORNADO	60,653	58,537,1
LIGHTNING	15,754	816	5,230	STORM SURGE	409	47,964,7
THUNDERSTORM	323,391	702	9,365	HAIL	288,661	15,732,2
RIP CURRENT	774	572	529	THUNDERSTORM	323,391	9,704,9
COLD	2,268	415	315	WILDFIRE	4,222	8,391,0
HIGH WIND	21,745	283	1,439	TROPICAL STORM	690	7,703,8
AVALANCHE	386	224	170	WINTER STORM	11,433	6,688,4
WINTER STORM	11,433	206	1,321	HIGH WIND	21,745	5,878,3
(OTHERS)	382,243	1,661	13,132	(OTHERS)	109,706	15,334,7
GRAND TOTAL	902,297	15,145	140,528	GRAND TOTAL	902,297	427,318,6

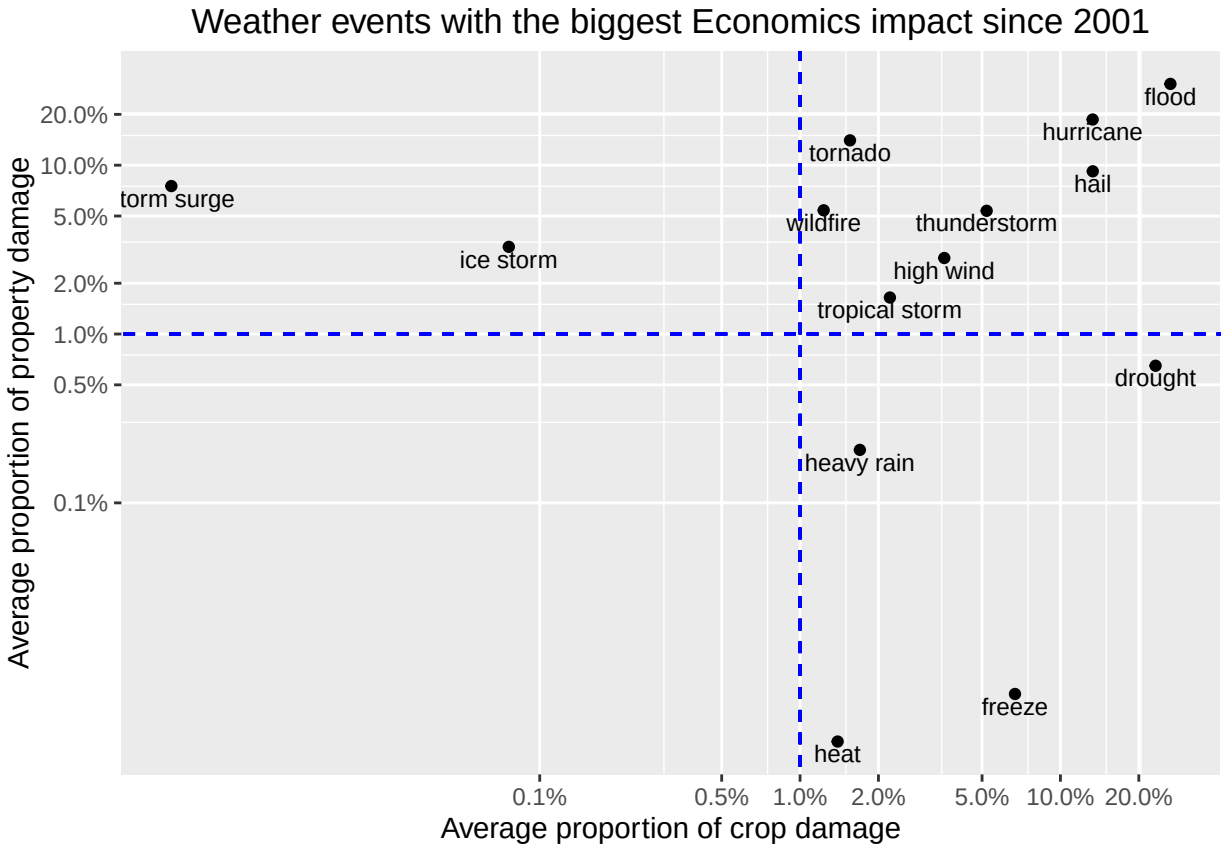
Results

```
first_year<-2001
lowest_percent=0.01

event_summary <- annual_events  %>% filter(year>first_year) %>% group_by(EVTYPE) %>%
                 summarise(number_of_events = sum(number_of_events),
                           event_injuries = sum(event_injuries),
                           event_fatalities = sum(event_fatalities),
                           event_property_damage = sum(event_property_damage),
                           event_crop_damage = sum(event_crop_damage),
                           m_fatality_percent = mean(event_fatality_percent),
                           m_injury_percent = mean(event_injury_percent),
                           m_property_percent = mean(event_property_damage_percent),
                           m_crop_percent = mean(event_crop_damage_percent))

par(mar=c(2,2,4,1))
ggplot(subset(event_summary, (m_fatality_percent >lowest_percent | m_injury_percent >lowest_percent)),
                aes(m_injury_percent, m_fatality_percent), height=800, width=800) +
                geom_point() +
                scale_y_continuous(labels=percent, breaks = c(0.001,0.01,0.02,0.03, 0.05,0.08, 0.1,0.15
                scale_x_continuous(labels=percent, breaks = c(0.005,0.01,0.02,0.03, 0.05, 0.1, 0.2, 0.4
                coord_trans(x="log10", y="log10") +
                geom_text(aes(label=tolower(EVTYPE)), hjust=0.5, vjust=1.2,
                          check_overlap = T, size=3) +
                ylab("Average proportion of annual fatalities") +
                xlab("Average proportion of annual injuries") +
```

```
              geom_segment(y=lowest_percent,yend=lowest_percent,x=0.0001,xend=1, colour="blue", linety
              geom_segment(y=0.0001,yend=1,x=lowest_percent,xend=lowest_percent, colour="blue", linety
              ggtitle(paste0("Weather events with the biggest impact on Population Health since ", fi
```




```
ggplot(subset(event_summary, m_crop_percent >lowest_percent | m_property_percent >lowest_percent),
              aes(m_crop_percent, m_property_percent), height=800, width=800) +
              geom_point() +
              scale_y_continuous(labels=percent, breaks = c(0.001, 0.005,0.01,0.02, 0.05, 0.1, 0.2, 0
              scale_x_continuous(labels=percent, breaks =  c(0.001,0.005, 0.01,0.02, 0.05, 0.1, 0.2, 
              coord_trans(x="log10", y="log10") +
              geom_text(aes(label=tolower(EVTYPE)), hjust=0.5, vjust=1.2,
                        check_overlap = F, size=3) +
              ylab("Average proportion of property damage") +
              xlab("Average proportion of crop damage") +
              geom_segment(y=lowest_percent,yend=lowest_percent,x=0.0000001,xend=1, colour="blue", li
              geom_segment(y=0.0000001,yend=1,x=lowest_percent,xend=lowest_percent, colour="blue", li
              ggtitle(paste0("Weather events with the biggest Economics impact since ", first_year))
```


Weather events with the biggest Economics impact since 2001
Average proportion of property damage
20.0%
10.0%
5.0%
2.0%
1.0%
0.5%
0.1%
flood
hurricane
tornado
hail
torm surge
wildfire
thunderstorm
ice storm
high wind
tropical storm
drought
heavy rain
freeze
heat
0.1%
0.5%
1.0%
2.0%
5.0%
10.0%
20.0%
Average proportion of crop damage