StormData Analysis Report: Fatality and Property Damage Caused by Weather Events Across USA as Revealed by NOAA Historical Weather Data

Ke Xu

February 29, 2016

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.

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.

The data set has 902297 records, collected from 1/1/1966 to 9/9/2011. We are using this data set to see the impact of weather events on human lives and property damages. since this is only for illustration purposes, I only did analysis on fatalities and property damages, skipped injury and crip damage data

Data Processing

Load data:

The data set is huge, including huge amount of free text remarks and lots of events that don't cause human or property damages. So a simplified data set, sdata, is created from original data set. Subsequenct analyses are performed on sdata . sdata only include a subset of columns and rows with fatality, injury or property/crop damage values.

```
library(plyr)
library(ggplot2)
rm(list=ls())

StormData <- read.csv ("repdata_data_StormData.csv")

sdata = StormData[c("BGN_DATE", "COUNTYNAME", "STATE", "EVTYPE", "FATALITIES",
"INJURIES",
  "PROPDMG", "PROPDMGEXP", "CROPDMG", "CROPDMGEXP", "REFNUM") ]
sdata = sdata [ ((sdata$PROPDMG>0)|(sdata$CROPDMG>0)|(sdata$FATALITIES>0)|(sdat
a$INJURIES>0)), ]
```

Data Cleaning

The following were done in data cleaning. (1) There is one flood in Napa county, California, showing $115 billion property damage (Refnum 605943). The remarks of this record noted the property damage was more than $70M. So I changed the exponent of this record to “M”. (2) Overall, the exponent columns are inconsistent. A new column PROPDMGMULT is created and used for calculating real damage values. (3) I tried to examine event types and clean them up, esepecially those more common types found in the data set. Some of the similar types are actually different. For example, “heat” and “excessive heat” are actually defined as two different types of events by NOAA. But “rip currents” and “rip current” are probably same and I merged these two. A new column, EVTYPE2, was created to store modified event types

```
# (1) correct error in record 605943
sdata$PROPDMGEXP [ (sdata$REFNUM==605943)] = "M"

# (2) calculate damage values
sdata$PROPDMGMULT = 0
unique(sdata$PROPDMGEXP)
```

```
##  [1] K M   B m + 0 5 6 4 h 2 7 3 H -
## Levels:  - ? + 0 1 2 3 4 5 6 7 8 B h H K m M
```

```
sdata$PROPDMGMULT [ sdata$PROPDMGEXP == "K"] = 1000
sdata$PROPDMGMULT [ sdata$PROPDMGEXP == "M"] = 1000000
sdata$PROPDMGMULT [ sdata$PROPDMGEXP == ""] = 1
sdata$PROPDMGMULT [ sdata$PROPDMGEXP == "B"] = 1000000000
sdata$PROPDMGMULT [ sdata$PROPDMGEXP == "m"] = 1000000
sdata$PROPDMGMULT [ sdata$PROPDMGEXP == "0"] = 1
sdata$PROPDMGMULT [ sdata$PROPDMGEXP == "5"] = 100000
sdata$PROPDMGMULT [ sdata$PROPDMGEXP == "6"] = 1000000
sdata$PROPDMGMULT [ sdata$PROPDMGEXP == "4"] = 10000
sdata$PROPDMGMULT [ sdata$PROPDMGEXP == "h"] = 100
sdata$PROPDMGMULT [ sdata$PROPDMGEXP == "2"] = 100
sdata$PROPDMGMULT [ sdata$PROPDMGEXP == "7"] = 10000000
sdata$PROPDMGMULT [ sdata$PROPDMGEXP == "3"] = 1000
sdata$PROPDMGMULT [ sdata$PROPDMGEXP == "H"] = 100
sdata$PROPDMGVAL = sdata$PROPDMG * sdata$PROPDMGMULT


sdata$CROPDMGMULT = 0
unique(sdata$CROPDMGEXP)
```

```
## [1]   M K m B ? 0 k
## Levels:  ? 0 2 B k K m M
```

```
sdata$CROPDMGMULT [ sdata$CROPDMGEXP == "M" ] = 1000000
sdata$CROPDMGMULT [ sdata$CROPDMGEXP == "K" ] = 1000
sdata$CROPDMGMULT [ sdata$CROPDMGEXP == "m" ] = 1000000
sdata$CROPDMGMULT [ sdata$CROPDMGEXP == "B" ] = 1000000000
sdata$CROPDMGMULT [ sdata$CROPDMGEXP == "0" ] = 1
sdata$CROPDMGMULT [ sdata$CROPDMGEXP == "k" ] = 1000
sdata$CROPDMGVAL = sdata$CROPDMG * sdata$CROPDMGMULT

# (3) clean up event types
evtypes = aggregate (x=sdata$EVTYPE, by=list(sdata$EVTYPE), FUN=NROW)
evtypes [ evtypes$x > 200, ]
```

```
##                  Group.1     x
## 11             AVALANCHE   268
## 14              BLIZZARD   253
## 49               DROUGHT   266
## 61        EXCESSIVE HEAT   698
## 73           FLASH FLOOD 20967
## 82        FLASH FLOODING   302
## 86                 FLOOD 10175
## 90     FLOOD/FLASH FLOOD   279
## 134                HAIL 26130
## 151                HEAT   215
## 159           HEAVY RAIN  1105
## 170           HEAVY SNOW  1342
## 200            HIGH WIND  5522
## 208           HIGH WINDS   657
## 238            ICE STORM   708
## 258            LIGHTNING 13293
## 306          RIP CURRENT   400
## 307         RIP CURRENTS   241
## 353          STRONG WIND  3370
## 364    THUNDERSTORM WIND 43655
## 381   THUNDERSTORM WINDS 12086
## 407              TORNADO 39944
## 417       TROPICAL STORM   416
## 423            TSTM WIND 63234
## 437       TSTM WIND/HAIL   441
## 456 URBAN/SML STREAM FLD   702
## 469     WILD/FOREST FIRE   388
## 471             WILDFIRE   857
## 481         WINTER STORM  1508
## 484       WINTER WEATHER   407
```

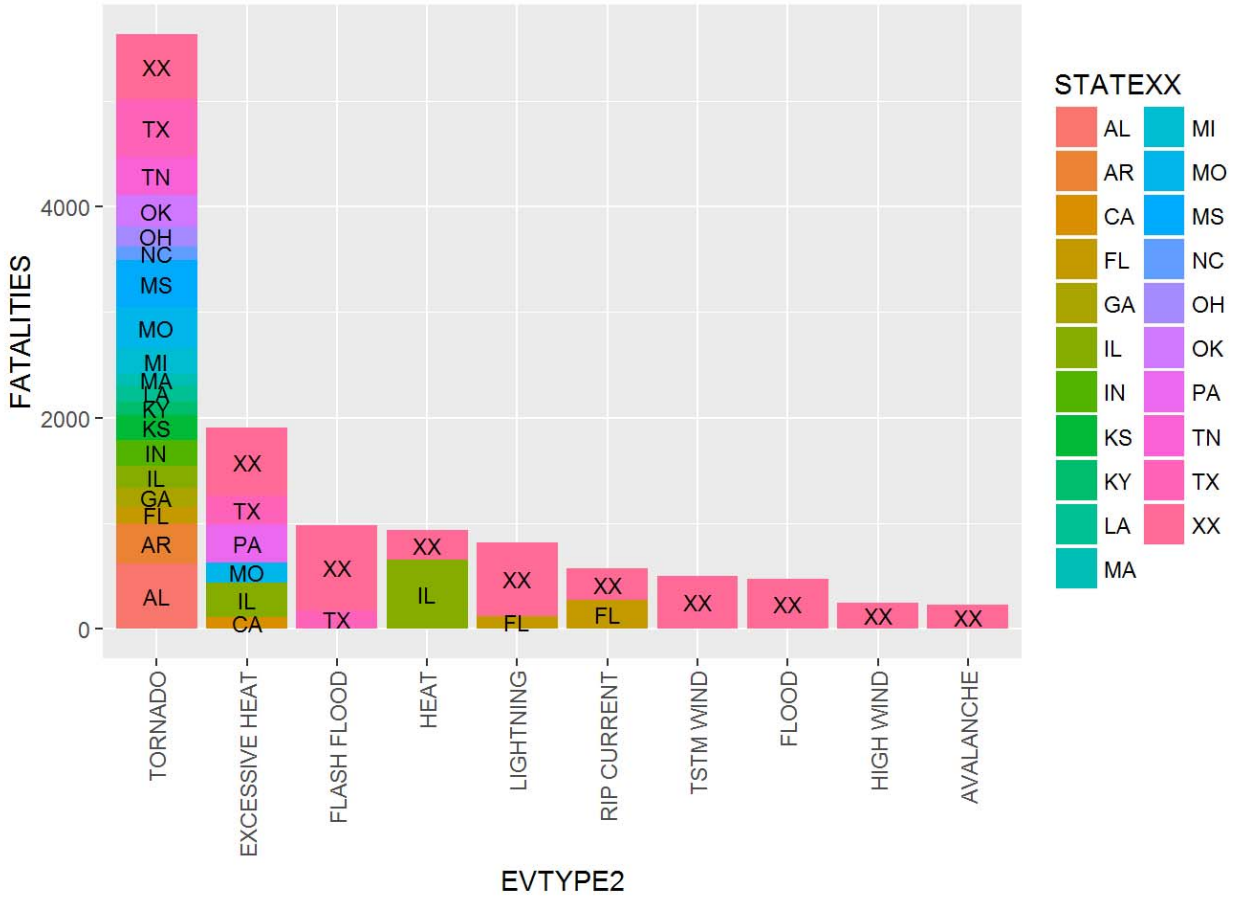
```
sdata$EVTYPE2 = sdata$EVTYPE
sdata$EVTYPE2 [ sdata$EVTYPE2 == "RIP CURRENTS"] = "RIP CURRENT"
```

Results

Fatalities

Here I am trying to show 10 weather events that caused most fatalities and states with most cases with each type of events. States with 100 or more fatalities are shown. Other states are grouped together in the plot

```
# fatal stores aggregated fatalities by events, fatal2 stores per state number
fatal <- aggregate(FATALITIES ~ EVTYPE2, data = sdata, FUN = sum)
fatal2 <- aggregate(FATALITIES ~ EVTYPE2 + STATE, data = sdata, FUN = sum)

# STATEXX created to state code or XX for "others"
# states with less than 100 deaths for any event are grouped together
fatal2$STATEXX = fatal2$STATE
fatal2$STATEXX [ fatal2$FATALITIES < 100 ] = "XX"

# re-aggregate state with new category "XX"
fatal2a <- aggregate(FATALITIES ~ EVTYPE2 + STATEXX, data = fatal2, FUN = sum)

# only selecting top 10 events
fatal3 = fatal2a [ ((fatal2a$EVTYPE2 == "TORNADO")|(fatal2a$EVTYPE2=="EXCESSIV
E HEAT")
  |(fatal2a$EVTYPE2=="FLASH FLOOD")|(fatal2a$EVTYPE2=="HEAT")|(fatal2a$EVTYPE2=
="LIGHTNING")
  |(fatal2a$EVTYPE2=="RIP CURRENT")|(fatal2a$EVTYPE2=="TSTM WIND")|(fatal2a$EVT
YPE2=="FLOOD")
  |(fatal2a$EVTYPE2=="HIGH WIND")|(fatal2a$EVTYPE2=="AVALANCHE")), ]
## generate position for displaying state names in bargraph
fatal4 <- ddply(fatal3, .(EVTYPE2),
               transform, pos = cumsum(FATALITIES) - (0.5 * FATALITIES)
)

# factor level for display purpose
fatal4$EVTYPE2 = as.vector(fatal4$EVTYPE2)
fatal4$EVTYPE2 = factor(fatal4$EVTYPE2, levels=c("TORNADO", "EXCESSIVE HEAT",
"FLASH FLOOD", "HEAT",
  "LIGHTNING", "RIP CURRENT", "TSTM WIND", "FLOOD", "HIGH WIND", "AVALANCHE"),
ordered=TRUE)

ggplot(fatal4, aes(x = EVTYPE2,  y = FATALITIES)) +
  geom_bar(aes(fill = STATEXX), stat="identity") +
  geom_text(aes(label = STATEXX, y = pos), size = 3) +
  theme(axis.text.x= element_text(angle=90,hjust=1,vjust=0.5))
```

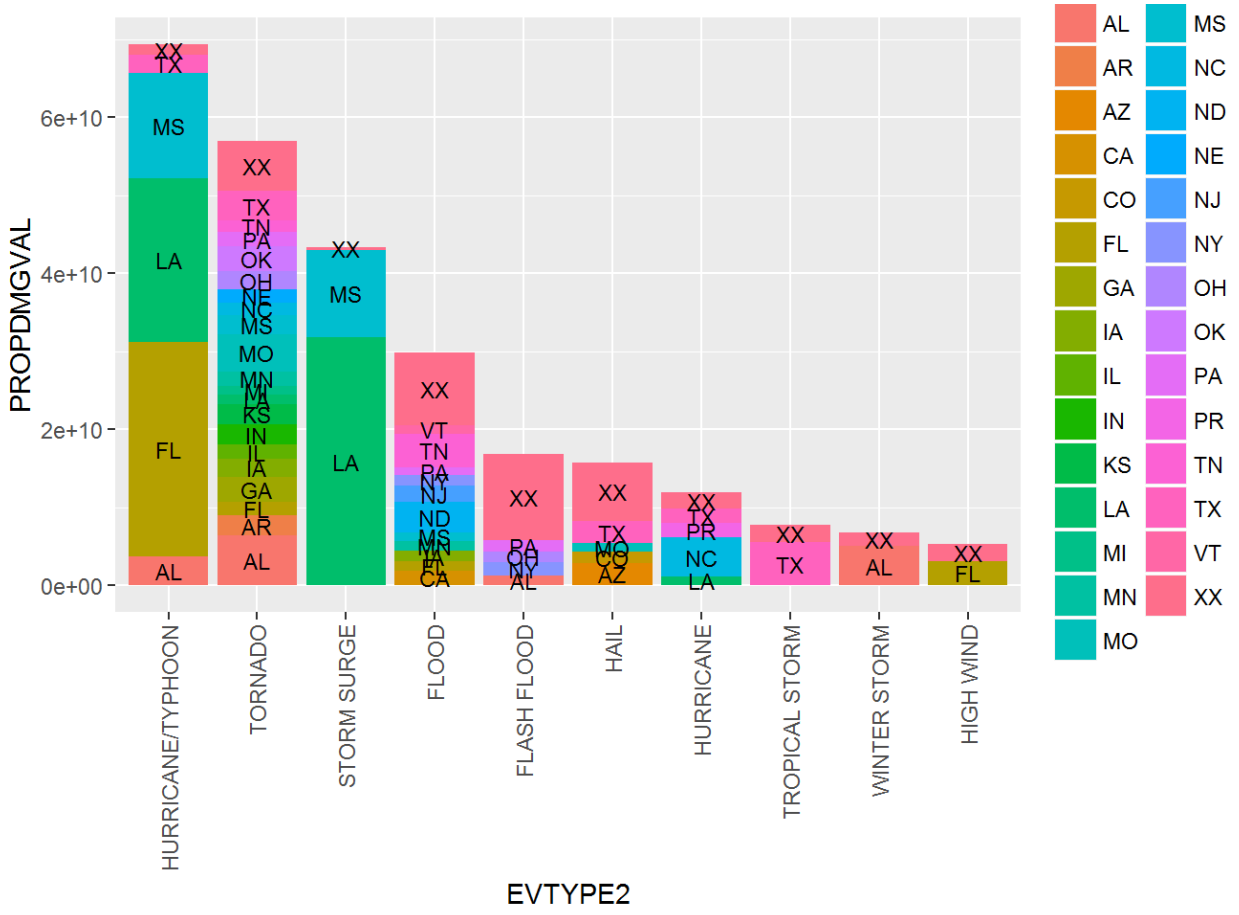


It is clear tornado-ally-states are responsible for most of the tornado deaths. It is also interesting that Illinois has most deaths from heat but not excessive heat (temperature not that high but more deaths due to some unknown reason?).

Property Damages

In this plot, states with accumulated damage more than one billion dollar are shown.

```
property <- aggregate(PROPDMGVAL ~ EVTYPE2, data = sdata, FUN = sum)
property2 <- aggregate(PROPDMGVAL ~ EVTYPE2 + STATE, data = sdata, FUN = sum)

# here we are grouping state with less than one Billion dollar damages
property2$STATEXX = property2$STATE
property2$STATEXX [ property2$PROPDMGVAL < 1000000000 ] = "XX"

property2a <- aggregate(PROPDMGVAL ~ EVTYPE2 + STATEXX, data = property2, FUN
= sum)

property3 = property2a [ ((property2a$EVTYPE2 == "HURRICANE/TYPHOON")|(property
2a$EVTYPE2=="TORNADO")
                     |(property2a$EVTYPE2=="STORM SURGE")|(property2a$EVTYPE2=
="FLOOD")|(property2a$EVTYPE2=="FLASH FLOOD")
                     |(property2a$EVTYPE2=="HAIL")|(property2a$EVTYPE2=="HURRICA
NE")|(property2a$EVTYPE2=="TROPICAL STORM")
                     |(property2a$EVTYPE2=="WINTER STORM")|(property2a$EVTYPE2=
="HIGH WIND")), ]
## generate position for displaying state names in bargraph
property4 <- ddply(property3, .(EVTYPE2),
                 transform, pos = cumsum(PROPDMGVAL) - (0.5 * PROPDMGVAL)
)


property4$EVTYPE2 = as.vector(property4$EVTYPE2)
property4$EVTYPE2 = factor(property4$EVTYPE2, levels=c("HURRICANE/TYPHOON", "TO
RNADO", "STORM SURGE", "FLOOD",
             "FLASH FLOOD", "HAIL", "HURRICANE", "TROPICAL STORM", "WINTER STOR
M", "HIGH WIND"), ordered=TRUE)

ggplot(property4, aes(x = EVTYPE2,  y = PROPDMGVAL)) +
  geom_bar(aes(fill = STATEXX), stat="identity") +
  geom_text(aes(label = STATEXX, y = pos), size = 3) +
  theme(axis.text.x= element_text(angle=90,hjust=1,vjust=0.5))
```



Storm surge is the 3rd largest event by property damage, with total damage amount around 43 Billion dollars. But almost all of the damages were actually from one event, Hurricane Katrina, which is recorded in two records, one for Louisiana (Ref#577616, propdmg 31B) and the other for Missisippi (Ref#581535, propdmg 11B)