Preliminary Data Analysis of Severe Weather Events in the USA

qpxu007

August 6, 2015

Synopsis

In the following excercise, we analyze the severe weather events collected by NOAA to answer two questions: 1. Across the United States, which types of events (as indicated in the EVTYPE variable) are most harmful with respect to population health? 2. Across the United States, which types of events have the greatest economic consequences?

Introduction

The severe weather event data collected by NOAA contains 902297 obs. of 37 variables. The "EVTYPE" column contains type of events. The health related columns "FATALITIES" and "INJURIES". The economic damage related columns are property damage ("PROPDMG" and "PROPDMGEXP") and crop damage("CROPDMG" and "CROPDMGEXP").

Data Processing

Firt, we read the data into a data frame (df):

```
df<-read.csv("repdata-data-StormData.csv")
str(df)
```

```
## '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
##  $ BGN_TIME  : Factor w/ 3608 levels "00:00:00 AM",..: 272 287 2705 1683 2584 3186 242 1683 3186 318
##  $ 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 1
##  $ 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 8
##  $ 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 "","" 
##  $ 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/ 436774 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 ...
```

To answer the questions raised above, we need to extract the following columns: "EVTYPE", "FATALITIES", "INJURIES", "PROPDMG", "PROPDMGEXP", "CROPDMG", and "CROPDMGEXP". For the damages, it should be noted that, the numbers in CROPDMG/PROPDMG columns are given in different units (kilo, mil, billion), which are set in the corresponding CROPDMGEXP/PROPDMGEXP columns. As a result, we need to do some cleaning up and conversion. So we subset part of the data into df2:

```
s<-c("EVTYPE","FATALITIES","INJURIES",
     "PROPDMG","PROPDMGEXP", "CROPDMG", "CROPDMGEXP")
df2 <- df[s]
```

For the damages, it should be noted that, the numbers in CROPDMG/PROPDMG columns are given in different units (kilo, mil, billion), which are stored in the corresponding CROPDMGEXP/PROPDMGEXP columns respectively. Below we convert all damages into the same unit (Millions, we only deal with H, K, M, and B, the rest are assumed to be 0 since they are small in comparison):

```
trans <-function(unit) {
  unit = toupper(unit)
  nfactor = 0.0
  if (unit == "K") {
    nfactor <- 0.001
  } else if (unit == "H") {
    nfactor <- 0.0001
  } else if (unit == "M") {
    nfactor <- 1.0
  } else if (unit == "B") {
    nfactor <- 1000.0
  }
  nfactor
}

df2$PROPDMGEXP<-sapply(df2$PROPDMGEXP, trans)
df2$PROPDMG<- df2$PROPDMG * df2$PROPDMGEXP

df2$CROPDMGEXP<-sapply(df2$CROPDMGEXP, trans)
df2$CROPDMG <- df2$CROPDMG * df2$CROPDMGEXP
```

We now can add two new columns: “health” to combine injuries and fatalities, and “damage” to combine crop and property damages. We can also clean up df2 a bit by removing the unit columns (“PROPDMGEXP”,“CROPDMGEXP”) that are no longer needed.

```
df2$health <- df2$INJURIES+df2$FATALITIES
df2$damage <- df2$PROPDMG+df2$CROPDMG
df2$PROPDMGEXP<- df2$CROPDMGEXP  <- NULL
str(df2)
```

```
## 'data.frame':    902297 obs. of  7 variables:
##  $ EVTYPE    : Factor w/ 985 levels "   HIGH SURF ADVISORY",..: 834 834 834 834 834 834 834 834 834 8
##  $ 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  0.025 0.0025 0.025 0.0025 0.0025 0.0025 0.0025 0.0025 0.025 0.025 ...
##  $ CROPDMG   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ health    : num  15 0 2 2 2 6 1 0 15 0 ...
##  $ damage    : num  0.025 0.0025 0.025 0.0025 0.0025 0.0025 0.0025 0.0025 0.025 0.025 ...
```

Results

To answer questions above, we first need to aggregate the df2 with *sum* by EVTYPE, to generate a new data frame df3.

```
df3 <- aggregate(df2[c("health","damage", "INJURIES",
                       "FATALITIES", "CROPDMG", "PROPDMG")],
                       by=list(EVTYPE=df2$EVTYPE), FUN=sum)
summary(df3)
```

```
##                    EVTYPE        health          damage
##     HIGH SURF ADVISORY:  1   Min.   :    0   Min.   :     0.00
##   COASTAL FLOOD       :  1   1st Qu.:    0   1st Qu.:     0.00
##   FLASH FLOOD         :  1   Median :    0   Median :     0.00
##   LIGHTNING           :  1   Mean   :  158   Mean   :   483.68
##   TSTM WIND           :  1   3rd Qu.:    0   3rd Qu.:     0.08
##   TSTM WIND (G45)     :  1   Max.   :96979   Max.   :150319.68
##  (Other)              :979
##     INJURIES         FATALITIES         CROPDMG
##  Min.   :    0.0   Min.   :   0.00   Min.   :    0.00
##  1st Qu.:    0.0   1st Qu.:   0.00   1st Qu.:    0.00
##  Median :    0.0   Median :   0.00   Median :    0.00
##  Mean   :  142.7   Mean   :  15.38   Mean   :   49.85
##  3rd Qu.:    0.0   3rd Qu.:   0.00   3rd Qu.:    0.00
##  Max.   :91346.0   Max.   :5633.00   Max.   :13972.57
##
##     PROPDMG
##  Min.   :     0.00
##  1st Qu.:     0.00
##  Median :     0.00
##  Mean   :   433.83
##  3rd Qu.:     0.05
##  Max.   :144657.71
##
```

The data frame df3 contains necessary information to answer the questions above. In the following, we create a function that process the data frame based on two columns (EVTYPE, and a consequence): first sort the data frame in decreasing order of the consequence; and then plot the top n EVTYPE (x) against consequence (y, in log scale and decreasing order).

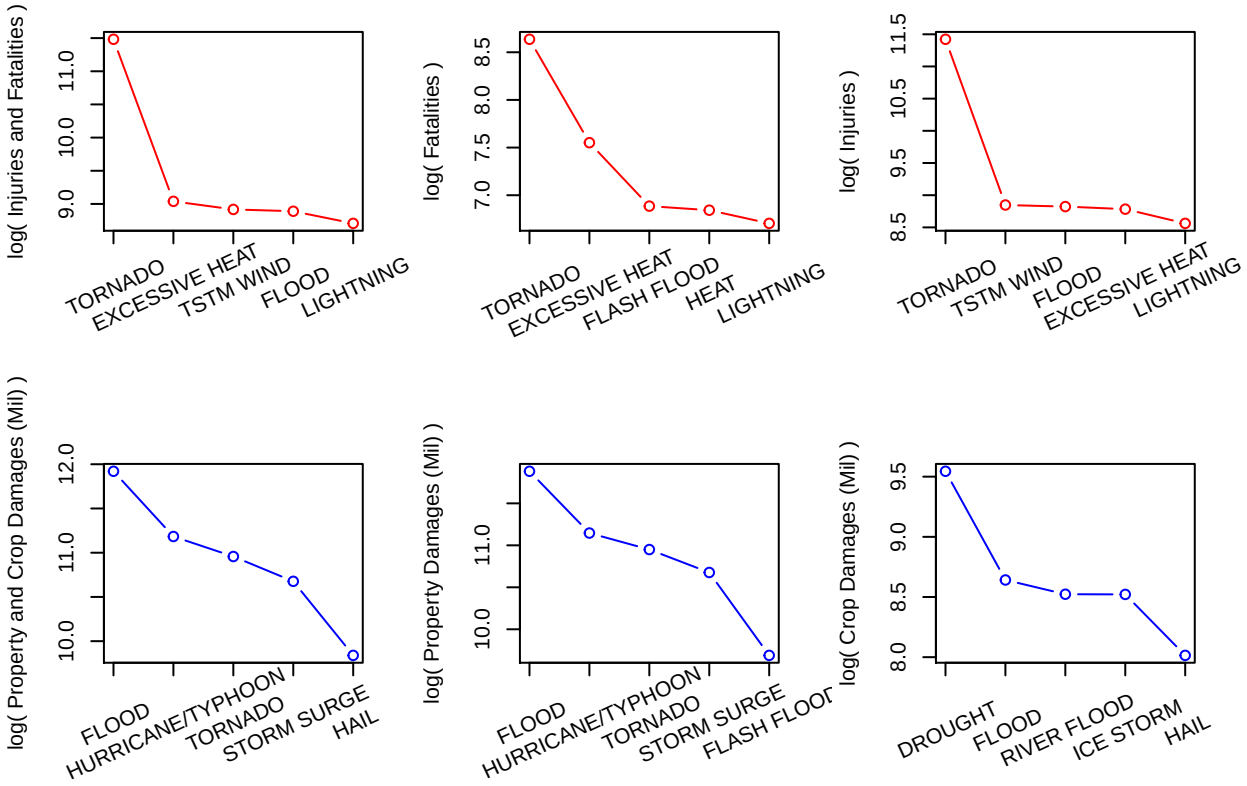
```
myplot<-function(df, xcol="EVTYPE", ycol="health", n=5,
                 ylabel="Injuries and Fatalities",
                 color="red", offset=1.5) {
   # sorted the df according the relevant ycol
   df <- df[order(-df[,ycol]), ]
   # select relevant columns and plot
   y <- df[1:n,ycol]
   x <- df[1:n,xcol]
   seqx <- seq(x)

   plot(seqx, log(y), type='b', xaxt='n',
        ylab=paste("log(",ylabel,")"), xlab="", col=color)
   axis(1,at=seqx, labels=F)
   text(seqx, par("usr")[3] - 0.2, labels = x, srt = 25, pos = 1,
        xpd = TRUE, offset=offset, cex=1)
}
```

Now, we can prepare a composite plots of EVTYPE vs various damages:

```
par(mfrow=c(2,3))
myplot(df3,ycol='health', ylabel="Injuries and Fatalities")
myplot(df3,ycol='FATALITIES', ylabel="Fatalities")
myplot(df3,ycol='INJURIES', ylabel="Injuries")

myplot(df3,ycol='damage', ylabel="Property and Crop Damages (Mil)", color='blue')
myplot(df3,ycol='PROPDMG', ylabel="Property Damages (Mil)", color='blue')
myplot(df3,ycol='CROPDMG', ylabel="Crop Damages (Mil)", color='blue')
```



Based on the plot above, it can be observed that tornados cause most fatalities and injuries, while floods cause most economical damages (property and crop combined, and property alone). Drougt is the leading cause of crop damage.

Discussion

Below we will explore mapping of the damage data to the maps. We can prep data similar to above analysis, and retain the location and date columns.

```
s<-c("EVTYPE","FATALITIES","INJURIES",
     "PROPDMG","PROPDMGEXP", "CROPDMG", "CROPDMGEXP",
     "STATE", "BGN_DATE", "LATITUDE", "LONGITUDE")
dfx <- df[s]

# fix date
dfx$BGN_DATE<-as.Date(as.character(dfx$BGN_DATE),
                      format="%m/%d/%Y %H:%M:%S")
# convert the lattitude and longitude data (assuming there are digital degrees)
dfx$LATITUDE<-dfx$LATITUDE*0.01
dfx$LONGITUDE<-dfx$LONGITUDE*0.01

dfx$PROPDMGEXP<-sapply(dfx$PROPDMGEXP,trans)
dfx$PROPDMG<- dfx$PROPDMG * dfx$PROPDMGEXP

dfx$CROPDMGEXP<-sapply(dfx$CROPDMGEXP,trans)
```

```
dfx$CROPDMG <- dfx$CROPDMG * dfx$CROPDMGEXP

dfx$health <- dfx$INJURIES+dfx$FATALITIES
dfx$damage <- dfx$PROPDMG+dfx$CROPDMG
dfx$PROPDMGEXP<- dfx$CROPDMGEXP  <- NULL

states<-cbind(tolower(state.name),state.abb)
colnames(states)<-c('region','STATE')
dfx<-merge(dfx, states, by=c("STATE"))
summary(dfx)
```

```
##      STATE                       EVTYPE         FATALITIES
##  TX     : 83728   HAIL              :288614   Min.   :  0.0000
##  KS     : 53440   TSTM WIND         :219807   1st Qu.:  0.0000
##  OK     : 46802   THUNDERSTORM WIND: 82477   Median :  0.0000
##  MO     : 35648   TORNADO           : 60635   Mean   :  0.0168
##  IA     : 31069   FLASH FLOOD       : 53274   3rd Qu.:  0.0000
##  NE     : 30271   FLOOD             : 24951   Max.   :583.0000
##  (Other):602228   (Other)           :153428
##     INJURIES            PROPDMG            CROPDMG
##  Min.   :   0.0000   Min.   :0.00e+00   Min.   :0.0e+00
##  1st Qu.:   0.0000   1st Qu.:0.00e+00   1st Qu.:0.0e+00
##  Median :   0.0000   Median :0.00e+00   Median :0.0e+00
##  Mean   :   0.1579   Mean   :4.80e-01   Mean   :5.5e-02
##  3rd Qu.:   0.0000   3rd Qu.:0.00e+00   3rd Qu.:0.0e+00
##  Max.   :1700.0000   Max.   :1.15e+05   Max.   :5.0e+03
##
##     BGN_DATE             LATITUDE       LONGITUDE          health
##  Min.   :1950-01-03   Min.   : 0.00   Min.   :  0.00   Min.   :   0.0000
##  1st Qu.:1995-01-07   1st Qu.:29.19   1st Qu.: 73.39   1st Qu.:   0.0000
##  Median :2001-09-07   Median :35.45   Median : 87.30   Median :   0.0000
##  Mean   :1998-11-01   Mean   :28.95   Mean   : 69.85   Mean   :   0.1747
##  3rd Qu.:2007-07-12   3rd Qu.:40.22   3rd Qu.: 96.14   3rd Qu.:   0.0000
##  Max.   :2011-11-30   Max.   :97.06   Max.   :166.12   Max.   :1742.0000
##
##      damage              region
##  Min.   :0.00e+00   texas    : 83728
##  1st Qu.:0.00e+00   kansas   : 53440
##  Median :0.00e+00   oklahoma : 46802
##  Mean   :5.30e-01   missouri : 35648
##  3rd Qu.:0.00e+00   iowa     : 31069
##  Max.   :1.15e+05   nebraska : 30271
##                     (Other) :602228
```

We can show the geospatial distribution of the human and economic costs of all the events for all time recorded.

```
df4<- aggregate(dfx[c("health","damage", "INJURIES",
                      "FATALITIES", "CROPDMG", "PROPDMG")],
                       by=list(region=dfx$region, STATE=dfx$STATE), FUN=sum)
```

We can use ggplot2 for this. The following charts divide the severity of the damage into five levels, and color them accordingly.

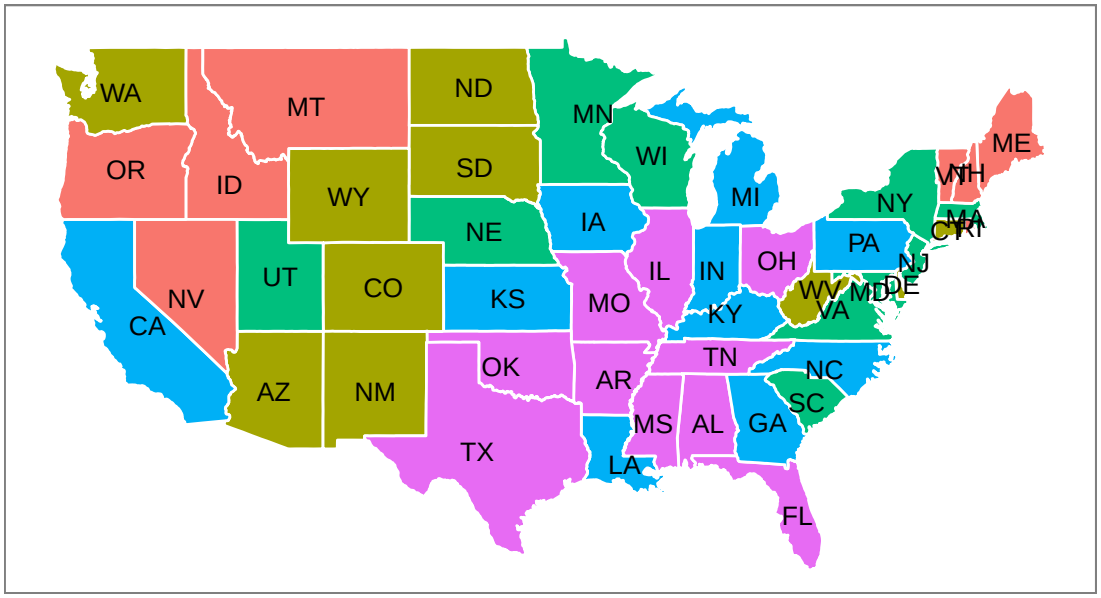
```
library(ggplot2)
library(ggthemes)
library(gridExtra)

states_map <- map_data("state")

#prepare state labels
cnames <-aggregate(cbind(long, lat) ~ region, data = states_map,
                   FUN = function(x) mean(range(x)))
states_full_abbrev <-cbind(tolower(state.name),state.abb)
colnames(states_full_abbrev)<-c('region','STATE')
cnames <-merge(cnames, states_full_abbrev, by=c('region'))
#tweak label positions
cnames[10, c(2:3)] <- c(-114.5, 43.5) # move label for idaho
cnames[16, 3] <- 30.6 #LA
cnames[20, c(2:3)] <- c(-84.5, 43) # MI
cnames[8, c(2:3)] <- c(-81.5, 28) # FL

title="Cumulative Injuries and Fatalities by State 1950-2011"
g1<-ggplot(df4, aes(map_id = region)) +
    geom_map(aes(fill = cut_number(health,5)), map = states_map, color ="white") +
    expand_limits(x = states_map$long, y = states_map$lat) +
    ggtitle(title) + theme_bw()+
    theme(plot.title = element_text(size=20, face="bold", vjust=2),
          panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
          legend.position = "bottom",
          legend.title=element_blank(),
          axis.ticks = element_blank(),
          axis.title = element_blank(),
          axis.text =  element_blank()) +
  geom_text(data=cnames, aes(long, lat, label = STATE, map_id =NULL), size=3.5) +
  coord_map()
g1
```

umulative Injuries and Fatalities by State 1950−20



[55,362] (362,1.19e+03] (1.19e+03,2.77e+03] (2.77e+03,5.46e+03] (5.46e+03,1.9e+

```
title="Cumulative Property and Crop Damage by State \n1950-2011"
g2<-ggplot(df4, aes(map_id = region)) +
    geom_map(aes(fill = cut_number(damage,5)), map = states_map, color ="white") +
    expand_limits(x = states_map$long, y = states_map$lat) +
    ggtitle(title) + theme_bw()+
    theme(plot.title = element_text(size=20, face="bold", vjust=2),
          panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
          legend.position = "bottom",
          legend.title=element_blank(),
          axis.ticks = element_blank(),
          axis.title = element_blank(),
          axis.text =  element_blank()) +
  geom_text(data=cnames, aes(long, lat, label = STATE, map_id =NULL), size=3.5) +
  coord_map()
g2
```

Cumulative Property and Crop Damage by State 1950−2011