Natural disasters in the US - An analysis

Mauro Leonelli

Tuesday, August 12, 2014

Synopsis

Natural disasters are causing damages for billion dollars and a huge cost in terms of human lifes and occur mostly unpredicted (and unpredictable). By analyzing the corpus of storm data from the U.S. National Oceanic and Atmospheric Administration's (NOAA) in the 1990 to 2011 period, it is possible nevertheless to identify areas and typologies of events to keep on focus. Property and crops are mostly afflicted by floods, which may be more sporadic but much more intense, while tornados and extreme heat have an higher cost in human lifes and tend to be much more stable over the different years. Floods must then be prevented or mitigated as much as possible especially in highly populated areas (California's flood of 2006 clearly shows the hugh cost in property damages), while tornados, affecting a huge area in Mid-West requires limited investments by means of a better tornado alert system which may mitigate the casualties.

Data Analysis

Pre-requisites In order to perform data analysis I made use of 2 popular libraries available in CRAN:

- plyr, a library to summarize and analyze data.
- gplot2, a powerful library for plotting data implementing the grammar of graphics.

```
library(plyr)
library(ggplot2)
```

In addition the Storm dataset needs to be downloaded in advance to avoid unnecessary data transfer.

Data Reading Dataset is in CSV bz2-compressed format and can be read with the standard read.csv command.

In order to get a more recent view at natural disasters in the US, I decided to limit the analysis from 1990 onwards; for this reason the dataset has been subset accordingly.

```
data <- read.csv("C:/R/RepData_PeerAssessment2/repdata_data_StormData.csv.bz2")
data$YEAR <- format(strptime(data$BGN_DATE, format = '%m/%d/%Y'), '%Y')
```

```
data2 <- data[data$YEAR >= 1990,]
```

Data Cleaning By mean of Storm documentation and dataset visual inspection, I determined a subset of observations that are vital for the data analysis in progress. Those observations are:

- BGN_DATE(character) : Date of the disaster occurence
- STATE (character) : 2-characters US State code
- EVTYPE(character): factor representing different natural events
- FATALITIES (numeric): number of deaths reported
- INJURIES (numeric): number of injured reported
- PROPDMG (numeric) : value of property damages (in PROPDMGEXP)

- PROPDMGEXP (factor): the unit of measure for PROPDMG (thousands, millions or billions)
- CROPDMG (numeric): value of crop damages (in CROPDMGEXP)
- CROPDMGEXP (factor): the unit of measure for CROPDMG (thousands, millions or billions)

```
data2 <- data2[, c("BGN_DATE", "STATE", "EVTYPE", "FATALITIES", "INJURIES", "PROPDMG", "PROPDMGEXP", "C
```

As described above, it is necessary to convert the EXP factors into their corresponding numeric values.

For performance considerations, the best technique is to create a support data.frame with the conversion values and apply them to our dataset with a left-join-style merge operations.

```
values = data.frame(c('K', 'M','B'), c(1000,1000000,1000000000))
names(values) = c("PROPDMGEXP", "PROPVALUE")
data2 <- merge(data2, values, by="PROPDMGEXP", all.x = TRUE)
names(values) = c("CROPDMGEXP", "CROPVALUE")
data2 <- merge(data2, values, by="CROPDMGEXP", all.x = TRUE)
data2[is.na(data2)] <- 0
```

Data Processing In order to fully analyze our dataset, the events have been aggregated by year and type. Affected counts the total number of fatalities and injuries per event, while damages sums both the property damages and crop damages, in billions of dollars.

```
dt <- ddply(data2, .(EVTYPE, YEAR), summarise, affected = sum(FATALITIES + INJURIES), damages = sum((((
```

From this dataset, I further aggregate data from the entire period, in order to get a summary view of the most dangerous events both by damage and population affected.

The top 10 overall most dangerous events both in terms of damage and people affected list, filtered to take unique factors only, is used to subset the original dataset for plotting purposes.

```
dt2 <- ddply(dt, .(EVTYPE), summarise, affected = sum(affected), damages = sum(damages))

uni <- unique(as.character(head(dt2[order(-dt2$affected),],10)$EVTYPE), as.character(head(dt2[order(-dt2

dt3 <- dt[dt$EVTYPE %in% uni,]
```

Finally, we use the most dangerous events to get a view of the US States most affected by them.

```
dt4 <- ddply(data2[data2$EVTYPE %in% uni, ], .(EVTYPE, STATE), summarise, affected = sum(FATALITIES + I
```

Results

First we check which are the most dangerous event types by affected people and damages.

```
  print(head(dt2[order(-dt2$affected),],10), type='html')
```

```
##                EVTYPE affected  damages
## 834           TORNADO    28426  30.8620
## 130    EXCESSIVE HEAT     8428   0.5002
## 170             FLOOD     7259 150.3197
## 464         LIGHTNING     6046   0.9408
```

```
## 856         TSTM WIND     5349   5.0389
## 275              HEAT     3037   0.4033
## 153        FLASH FLOOD     2755  17.5621
## 427         ICE STORM     2064   8.9670
## 760 THUNDERSTORM WIND     1621   3.8980
## 972      WINTER STORM     1527   6.7154
```

```r
print(head(dt2[order(-dt2$damages),],10), type='html')
```

```
##                EVTYPE affected damages
## 170             FLOOD     7259 150.320
## 411 HURRICANE/TYPHOON     1339  71.914
## 670       STORM SURGE       51  43.324
## 834           TORNADO    28426  30.862
## 244              HAIL     1154  18.753
## 153       FLASH FLOOD     2755  17.562
## 95            DROUGHT        4  15.019
## 402         HURRICANE      107  14.610
## 590       RIVER FLOOD        4  10.148
## 427         ICE STORM     2064   8.967
```

As it can be recognize, Tornados and Floods have the worst impact overall, with a clear superiority in injuries and fatalities for the former and property and crop damages for the latter.

Succesively, I would like to understand which is the historical behavior of natural disasters, both in terms of damages and affected people.

Let's look first at affected people.

```r
g <- ggplot(aes(YEAR, affected), data = dt3, colour=EVTYPE) + ylab("Total fatalities") + geom_point(aes

print(g)
```

Here the same plot for damages.

```r
g2 <- ggplot(aes(YEAR, damages), data = dt3, colour=EVTYPE) + ylab("Damages (in M$)")+ geom_point(aes(g

print(g2)
```

The graphs show:

- A general tendency for tornados and heat to affect regurarly population all along the 1990 - 2011 period
- Floods are not regularly but have a high count in injuries and damage when happening
- All the other natural events have a relatively small impact in term of damages

Last analysis is understanding which states are most affected by events

```r
print(head(dt4[order(-dt4$affected),],10), type='html')
```

```
##            EVTYPE STATE affected  damages
## 151         FLOOD    TX     6387 0.975417
## 362       TORNADO    AL     4261 5.165207
```

```
## 24  EXCESSIVE HEAT    MO     3715 0.000379
## 404        TORNADO    TN     2567 1.368240
## 385        TORNADO    MO     2439 3.841407
## 397        TORNADO    OK     2018 1.904620
## 371        TORNADO    GA     1919 1.291779
## 237      ICE STORM    OH     1654 0.207891
## 363        TORNADO    AR     1564 1.666609
## 386        TORNADO    MS     1418 1.610882
```

```
print(head(dt4[order(-dt4$damages),],10), type='html')
```

```
##           EVTYPE STATE affected damages
## 110        FLOOD    CA       48 117.378
## 362      TORNADO    AL     4261   5.165
## 227    ICE STORM    MS        0   5.025
## 468 WINTER STORM    AL        8   5.002
## 150        FLOOD    TN       36   4.250
## 135        FLOOD    ND       10   3.990
## 385      TORNADO    MO     2439   3.841
## 119        FLOOD    IA      122   2.970
## 138        FLOOD    NJ       25   2.112
## 397      TORNADO    OK     2018   1.905
```