Study of Injuries Caused by Several Natural Disasters

Javier Santibañez

Thursday, December 18, 2014

Synopsis

In this work we explore the injuries caused by several types of natural disasters. First we have to get and process the data, then we have to analize it in order to answer these questions:

1.- Which types of events (as indicated in the EVTYPE variable) are most harmful with respect to population health?

2.- Which types of events have the greatest economic consequences?

Data Processing

We use for the anályis the following packages:

- `R.utils`, to unzip the data file.
- `dplyr`, to handle the data set on R.
- `lurbitade`, to handle with dates and times.
- `ggplot2`, to make plots.

```
suppressMessages(suppressWarnings(require("R.utils")))
suppressMessages(suppressWarnings(require("dplyr")))
suppressMessages(suppressWarnings(require("lubridate")))
suppressMessages(suppressWarnings(require("ggplot2")))
```

First we have to download the data from the web and then load the data into R.

```
setwd("C:/Users/demyc 13/Documents/Course_Project_2")
url<-'http://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2'
zip_file<-'StormData.csv.bz2'
csv_file<-'StormData.csv'
if (!file.exists(zip_file)) download.file(url,zip_file)
if (!file.exists(csv_file)) bunzip2(zip_file,destname=csv_file,remove=F)
data<-tbl_df(read.csv(csv_file,stringsAsFactors=F))
```

Now, we have to select the relevant variables for this study, which are: type of disaster, date of occurency, and injuries. Variables were transformed into convenient formats.

```
data1<-select(data,BGN_DATE,EVTYPE,FATALITIES,
              INJURIES,PROPDMG,PROPDMGEXP,CROPDMG,CROPDMGEXP) %>%
       mutate(BGN_DATE=mdy(gsub(" 0:00:00","",BGN_DATE)),
              EVTYPE=tolower(EVTYPE))
```

It is important to know that there are a lot of disasters which are redundant, for example in some cases there are typos. Hence, the next code xoes an exhaustive depuration of the variable `EVTYPE`. I do not show the whole code because it is to long, so you can see it on my github repository for this assignment.

After the depuration we have only 15 types of natural disasters and one unknown cathegory. We can see now the number of ocurrences for every type of disaster.

```
table(data1$EVTYPE)
```

```
##
##          cold contamination       drought          fire         flood
##        322896            21          2724          4239         86127
##           fog          heat     hurricane     landslide       oceanic
##          1835          2768           298           658          6538
##          rain         storm       tornado       unknown       volcano
##         11926        367701         67848           316            29
##          wind
##         26373
```

If we explore the variables `PROPDMG` and `CROPDMG` we find no errors.

```
summary(data1$PROPDMG)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       0       0      12       0    5000
```

```
summary(data1$CROPDMG)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     0.0     0.0     1.5     0.0   990.0
```

But when we look at the variables `PROPDMGEXP` and `CROPDMGEXP` we can find a lot of errors.

```
summary(factor(data1$PROPDMGEXP))
```

```
##             -      ?      +      0      1      2      3      4      5
## 465934      1      8      5    216     25     13      4      4     28
##      6      7      8      B      h      H      K      m      M
##      4      5      1     40      1      6 424665      7  11330
```

```
summary(factor(data1$CROPDMGEXP))
```

```
##             ?      0      2      B      k      K      m      M
## 618413      7     19      1      9     21 281832      1   1994
```

So, we have to process this variable to depure errors. we are going to do this simultaneously for both variables.

```
data1$PROPDMGEXP<-tolower(data1$PROPDMGEXP)
data1$CROPDMGEXP<-tolower(data1$CROPDMGEXP)
data1$PROPDMGEXP[data1$PROPDMGEXP==""]<-0
data1$PROPDMGEXP[data1$PROPDMGEXP=="b"]<-9
data1$PROPDMGEXP[data1$PROPDMGEXP=="h"]<-2
data1$PROPDMGEXP[data1$PROPDMGEXP=="k"]<-3
data1$PROPDMGEXP[data1$PROPDMGEXP=="m"]<-6
data1$PROPDMGEXP[data1$PROPDMGEXP %in% c("-","?","+")]<-NA

data1$CROPDMGEXP[data1$CROPDMGEXP==""]<-0
data1$CROPDMGEXP[data1$CROPDMGEXP=="b"]<-9
data1$CROPDMGEXP[data1$CROPDMGEXP=="k"]<-3
data1$CROPDMGEXP[data1$CROPDMGEXP=="m"]<-6
data1$CROPDMGEXP[data1$CROPDMGEXP=="?"]<-NA
```

Now we can find no errors.

```
summary(factor(data1$PROPDMGEXP))
```

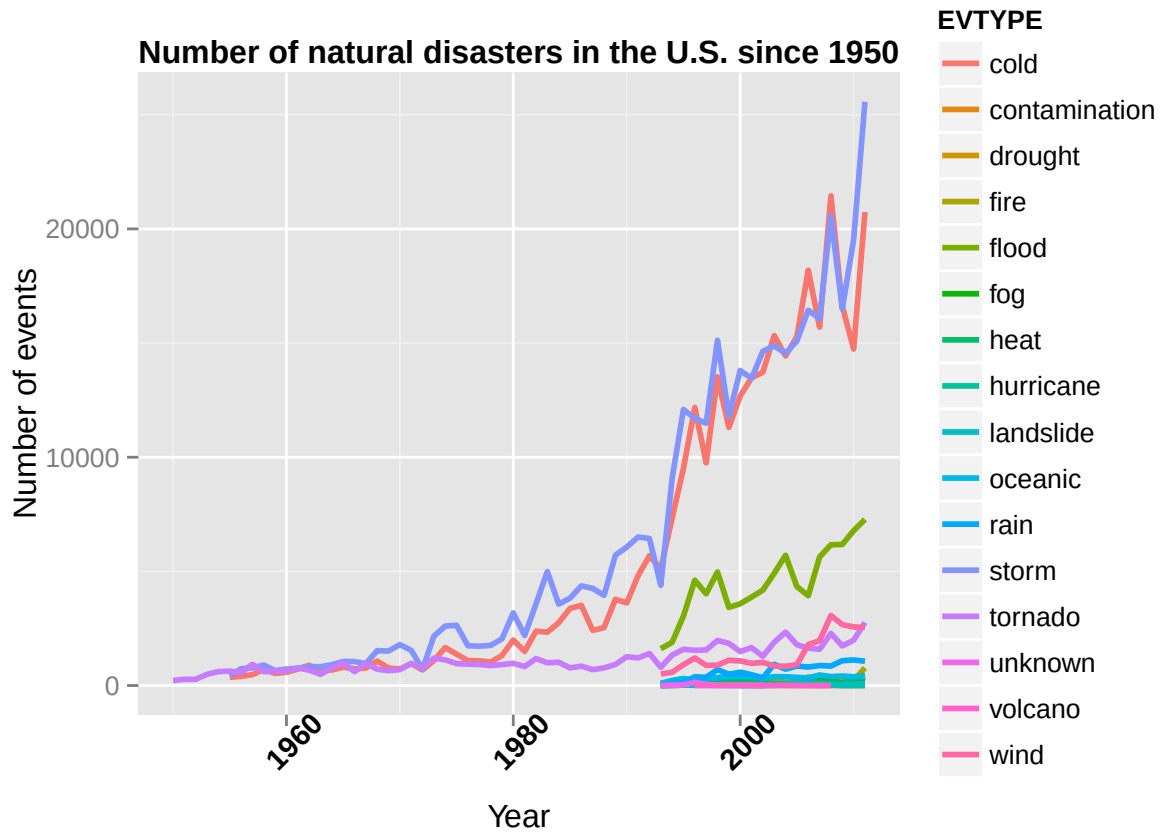
```
##      0      1      2      3      4      5      6      7      8      9
## 466150     25     20 424669      4     28  11341      5      1     40
##   NA's
##     14
```

```
summary(factor(data1$CROPDMGEXP))
```

```
##      0      2      3      6      9   NA's
## 618432      1 281853   1995      9      7
```

Finally, as many Coursera students have reported, there are not registers of all event types before 1993, The next graphic will help us to understand it.

```
plot0<-group_by(data1,EVTYPE,Year=year(BGN_DATE)) %>%
       summarise(n=n())
ggplot(plot0,aes(Year,n))+geom_line(aes(color=EVTYPE),size=1)+
  labs(title="Number of natural disasters in the U.S. since 1950")+
  xlab("Year")+ylab("Number of events")+
  theme(plot.title=element_text(size=12,face="bold"),
        axis.text.x=element_text(size=11,colour="black",face="bold",angle=45))
```



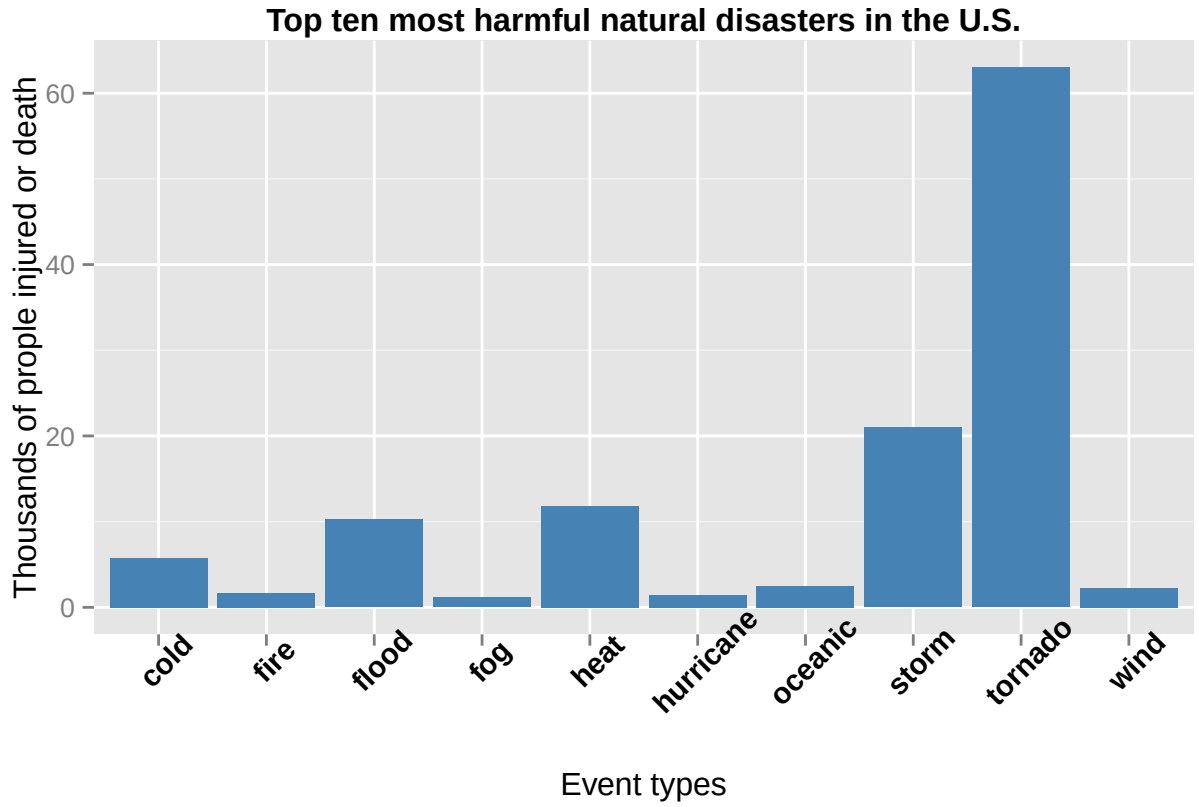
We can see that there are only registers of tornados, cold events and storms before 1993. So this does not mean that before 1993 there ware not other types of events in the U.S. but there are no registers. This fact will affect our next analysis, then we will subset the original data set keeping only information after 1993.

```
data1<-filter(data1,BGN_DATE>years(1993))
```

Results

We have to answer two questions. First question is about people health injuries. We are going to show a barplot with the number of deaths and/or injuries caused by the most dangerous natural disasters.

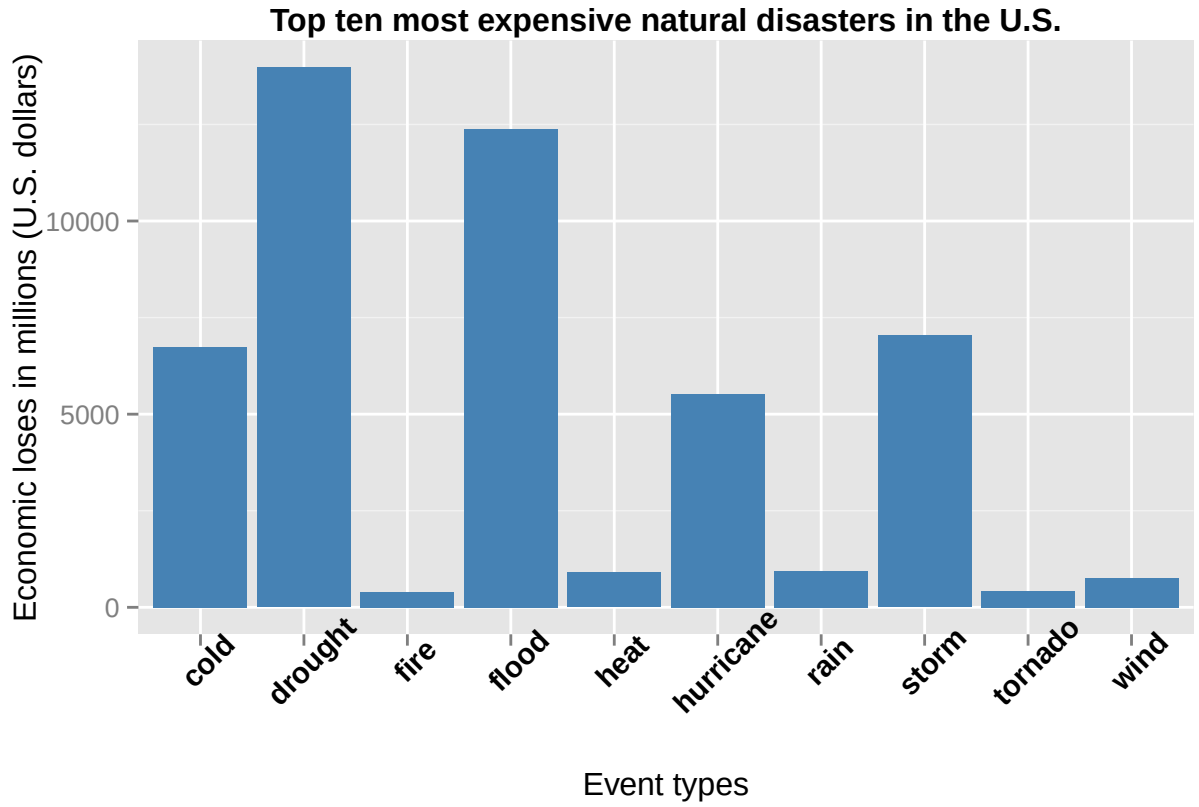
```
plot1<-group_by(data1,EVTYPE) %>%
       summarise(health=sum(FATALITIES,INJURIES,na.rm=T)/1000) %>%
       arrange(desc(health))
plot1<-plot1[1:10,]
ggplot(plot1,aes(EVTYPE,health))+
  geom_bar(stat="identity",fill="steelblue")+
  labs(title="Top ten most harmful natural disasters in the U.S.")+
  xlab("Event types")+ylab("Thousands of prople injured or death")+
  theme(plot.title=element_text(size=12,face="bold"),
        axis.text.x=element_text(size=11,colour="black",face="bold",angle=45))
```



Second question is about what natural disasters have the most greater economical consecuences. So, we have to compute the economic damage for every type of disaster.

```
data1$PROPDMG<-as.numeric(data1$PROPDMG)*10^as.numeric(data1$PROPDMGEXP)
data1$PROPDMG<-as.numeric(data1$CROPDMG)*10^as.numeric(data1$CROPDMGEXP)

plot2<-group_by(data1,EVTYPE) %>%
       summarise(econ.dmg=sum(PROPDMG,CROPDMG,na.rm=T)/10^6) %>%
       arrange(desc(econ.dmg))
plot2<-plot2[1:10,]
ggplot(plot2,aes(EVTYPE,econ.dmg))+
  geom_bar(stat="identity",fill="steelblue")+
  labs(title="Top ten most expensive natural disasters in the U.S.")+
  xlab("Event types")+ylab("Economic loses in millions (U.S. dollars)")+
  theme(plot.title=element_text(size=12,face="bold"),
        axis.text.x=element_text(size=11,colour="black",face="bold",angle=45))
```



Conclussion

Remember, we have to answer two questions:

- Which types of events are most harmful with respect to population health?
- Which types of events have the greatest economic consequences?

From our plot *Top ten most harmful natural disasters in the U.S.*, we can see that tornadoes have the first places, followed far behind by storms, heats, floods and colds.

On the other hand, from our plot *Top ten most expensive natural disasters in the U.S.* we can see that drought have the greatest economic consequences. Other disasters with important economic consequences are floods, storms, colds and hurricanes.