Study of the impact of weather event types in the USA

Jean-Luc BELLIER

19 aoÃ»t 2016

The goal of this project is to study the influence of some weather events on the USA population. This study will help in getting a better comprehension of the major weather events and their impact :

- on the population : how many hurst and deaths they cause
- on a financial matter (on the properties and crops)

The data are provided by the U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database. The data analysis will be done in two steps. First, the data are read and briefly summarized to retrieve the necessary information. In a second step, we will transform the data in a convenient manner, to be able to answer the desired questions.

Data Processing

The data are loaded directly from the compressed file. For efficiency reasons, we will put the data frame in cache. We will do a first raw analysis by looking at the column names, to identify which columns will be used for the analysis. We will also print the summary of these columns. The graphs will be plotted with the gglot2 library.

```r
library(R.cache)
library(dplyr)
```

```r
key<-list("0001")

stormData <- loadCache(key=key)
if (!is.null(stormData)) {
    print("Load cached data")
} else {
    stormData <- read.table("repdata_data_StormData.csv.bz2", header=T, quote="\"", sep=",")
    saveCache(stormData,key=key,comment="Load StormData")
}
```

```
## [1] "Load cached data"
```

```r
stormData_tbl_df <- tbl_df(stormData)
```

First analysis

The summary and the view of the column names help us in identifying the variables to keep for the analysis, and potential operations to do before getting further in the analysis. For practical use, the original data frame is transformed with the tbl_df function.

```r
names(stormData_tbl_df)
```

```
##  [1] "STATE__"    "BGN_DATE"   "BGN_TIME"   "TIME_ZONE"  "COUNTY"    
##  [6] "COUNTYNAME" "STATE"      "EVTYPE"     "BGN_RANGE"  "BGN_AZI"   
## [11] "BGN_LOCATI" "END_DATE"   "END_TIME"   "COUNTY_END" "COUNTYENDN"
## [16] "END_RANGE"  "END_AZI"    "END_LOCATI" "LENGTH"     "WIDTH"     
## [21] "F"          "MAG"        "FATALITIES" "INJURIES"   "PROPDMG"   
## [26] "PROPDMGEXP" "CROPDMG"    "CROPDMGEXP" "WFO"        "STATEOFFIC"
## [31] "ZONENAMES"  "LATITUDE"   "LONGITUDE"  "LATITUDE_E" "LONGITUDE_"
## [36] "REMARKS"    "REFNUM"
```

To answer the two questions, We will keep the following variables :

- EVTYPE : label of event type
- FATALITIES : number of deaths
- INJURIES : number of hurts
- CROPDMGEXP : code of unit for the amount of the damage on crops
- CROPDMG : amount of the damage on crops (in CROPDMGEXP unit)
- PROPDMGEXP : code of unit for the amount of the damage on properties
- PROPDMG : amount of the damage on properties (in PROPDMGEXP unit)

```
stormData1 <- stormData_tbl_df %>% select(EVTYPE, FATALITIES, INJURIES,CROPDMG, CROPDMGEXP, PROPDMG, PRO
summary(stormData1)
```

```
##                EVTYPE         FATALITIES         INJURIES     
##  HAIL             :288661   Min.   :  0.0000   Min.   :   0.0000  
##  TSTM WIND        :219940   1st Qu.:  0.0000   1st Qu.:   0.0000  
##  THUNDERSTORM WIND: 82563   Median :  0.0000   Median :   0.0000  
##  TORNADO          : 60652   Mean   :  0.0168   Mean   :   0.1557  
##  FLASH FLOOD      : 54277   3rd Qu.:  0.0000   3rd Qu.:   0.0000  
##  FLOOD            : 25326   Max.   :583.0000   Max.   :1700.0000  
##  (Other)          :170878                                         
##     CROPDMG          CROPDMGEXP        PROPDMG          PROPDMGEXP    
##  Min.   :  0.000          :618413   Min.   :   0.00          :465934  
##  1st Qu.:  0.000   K      :281832   1st Qu.:   0.00   K      :424665  
##  Median :  0.000   M      :  1994   Median :   0.00   M      : 11330  
##  Mean   :  1.527   k      :    21   Mean   :  12.06   0      :   216  
##  3rd Qu.:  0.000   0      :    19   3rd Qu.:   0.50   B      :    40  
##  Max.   :990.000   B      :     9   Max.   :5000.00   5      :    28  
##                    (Other):     9                     (Other):    84  
```

```
dim(stormData1)
```

```
## [1] 902297      7
```

Further analysis

We can see that all the values of the variables CROPDMG and PROPDMG are set up, because the summary does not show any NA value. Concerning the variables PROPDMGEXP and CROPDMGEXP, the number of values is low. These variables define the unit (i.e. the multiplicator) for the variables PROPDMG and CROPDMG.

- "B" for "Billion"

- “M” for “Million”
- “K” or “k” for “Kilo”
- “0” for standard unit

We will consider that all other values are standard (null string values, or “Other values”). The following function will help us in computing the multiplication factor. This can be done with the following function :

```
decodeUnit <- function(Unit)
{
    Decoded <- 1 # default value
    Decoded <- ifelse(Unit == "B",1000000000,Decoded)
    Decoded <- ifelse(Unit == "M",1000000,Decoded)
    Decoded <- ifelse(toupper(Unit) == "K",1000,Decoded)

    return(Decoded)
}
```

We compute the number of injuries and fatalities by event type, then sort the data by each aggregated variable separately :

Order number of injuries by event type :

```
## # A tibble: 985 × 3
##               EVTYPE sumInjuries sumFatalities
##               <fctr>       <dbl>         <dbl>
## 1            TORNADO       91346          5633
## 2          TSTM WIND        6957           504
## 3              FLOOD        6789           470
## 4     EXCESSIVE HEAT        6525          1903
## 5          LIGHTNING        5230           816
## 6               HEAT        2100           937
## 7          ICE STORM        1975            89
## 8        FLASH FLOOD        1777           978
## 9  THUNDERSTORM WIND        1488           133
## 10              HAIL        1361            15
## # ... with 975 more rows
```

Order number of fatalities by event type :

```
## # A tibble: 985 × 3
##            EVTYPE sumInjuries sumFatalities
##            <fctr>       <dbl>         <dbl>
## 1         TORNADO       91346          5633
## 2  EXCESSIVE HEAT        6525          1903
## 3     FLASH FLOOD        1777           978
## 4            HEAT        2100           937
## 5       LIGHTNING        5230           816
## 6       TSTM WIND        6957           504
## 7           FLOOD        6789           470
## 8     RIP CURRENT         232           368
## 9       HIGH WIND        1137           248
## 10      AVALANCHE         170           224
## # ... with 975 more rows
```

For the economical impacts, we will use the same kind of transformation, using the decoding function defined above. TOT_AMTPROPDMG is the total amount on damages on properties, and TOT_AMTCROPDMG the total amount of damages on crops. Both are in dollars.

```
## # A tibble: 10 × 3
##               EVTYPE TOT_AMTCROPDMG TOT_AMTPROPDMG
##               <fctr>          <dbl>          <dbl>
## 1            DROUGHT    13972566000     1046106000
## 2              FLOOD     5661968450   144657709807
## 3        RIVER FLOOD     5029459000     5118945500
## 4          ICE STORM     5022113500     3944927860
## 5               HAIL     3025954473    15727367053
## 6          HURRICANE     2741910000    11868319010
## 7  HURRICANE/TYPHOON     2607872800    69305840000
## 8        FLASH FLOOD     1421317100    16140812067
## 9       EXTREME COLD     1292973000       67737400
## 10      FROST/FREEZE     1094086000        9480000
```

```
## # A tibble: 10 × 3
##               EVTYPE TOT_AMTCROPDMG TOT_AMTPROPDMG
##               <fctr>          <dbl>          <dbl>
## 1              FLOOD     5661968450   144657709807
## 2  HURRICANE/TYPHOON     2607872800    69305840000
## 3            TORNADO      414953270    56925660790
## 4        STORM SURGE           5000    43323536000
## 5        FLASH FLOOD     1421317100    16140812067
## 6               HAIL     3025954473    15727367053
## 7          HURRICANE     2741910000    11868319010
## 8     TROPICAL STORM      678346000     7703890550
## 9       WINTER STORM       26944000     6688497251
## 10         HIGH WIND      638571300     5270046295
```

Results

Question 1 : Across the United States, which types of events (as indicated in the EVTYPE variable) are most harmful with respect to population health?

The following graph give the impact of the 10 major event types on the population. For the first question, we will use a line plot. To do this, we will add another column to distinguish the values between the number of injuries and the number of fatalities. This column will be used as a grouping column and will be used for the graph legend.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

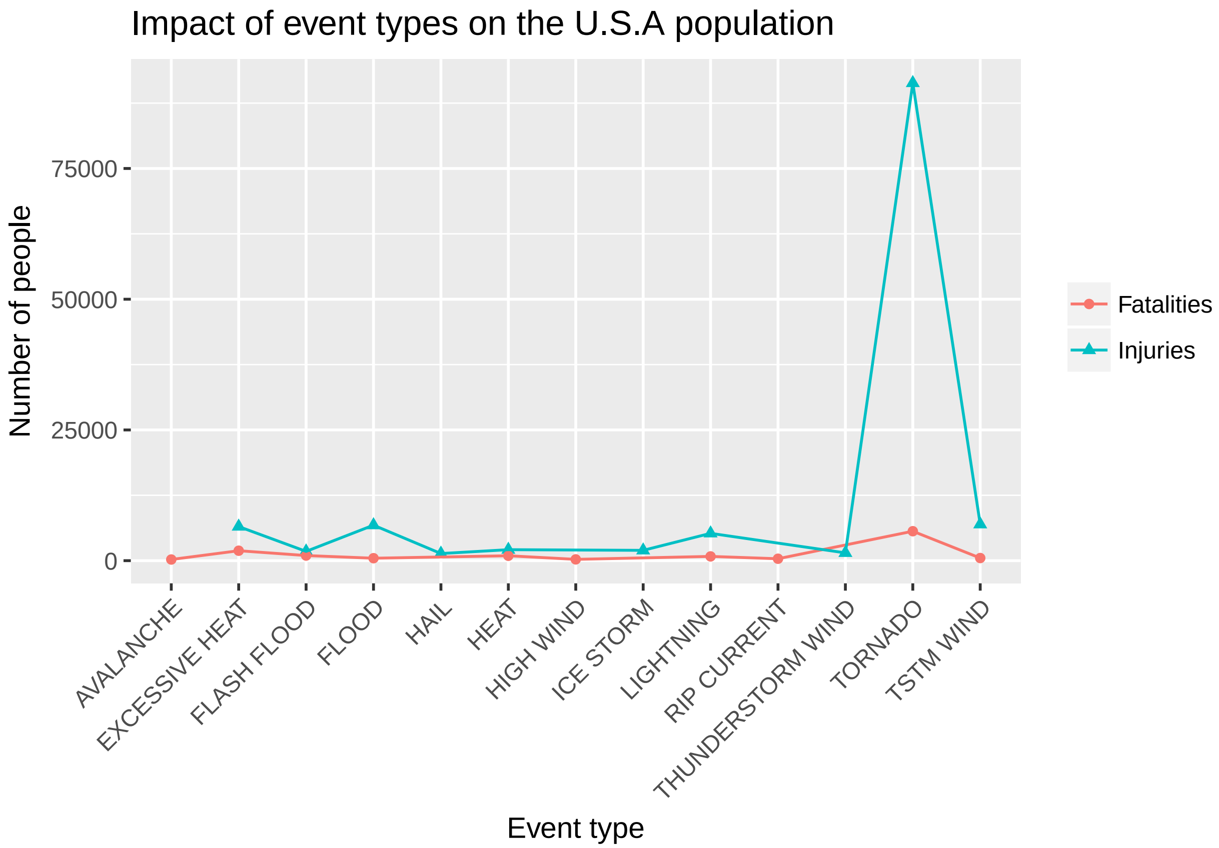
```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(reshape2)
```

```
## Warning: package 'reshape2' was built under R version 3.3.2
```

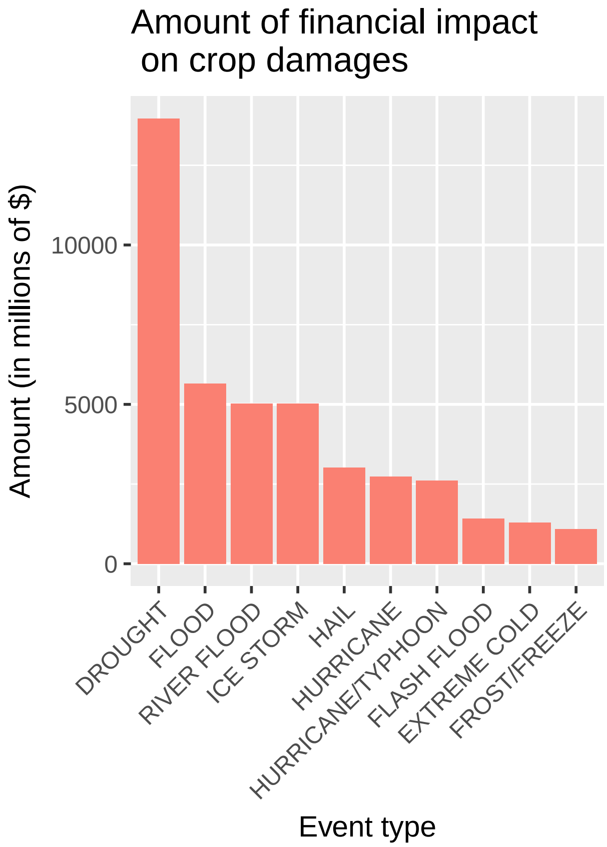
```
Impact1 <- mutate(select(stormDataMaxInjuries[1:10, ],EVTYPE,columnValue = sumInjuries),columnName ="In
Impact2 <- mutate(select(stormDataMaxFatalities[1:10, ],EVTYPE,columnValue = sumFatalities),columnName =

ggplot(data=rbind(Impact1,Impact2),aes(x=EVTYPE, y=columnValue, group=columnName,shape=columnName, colo
```

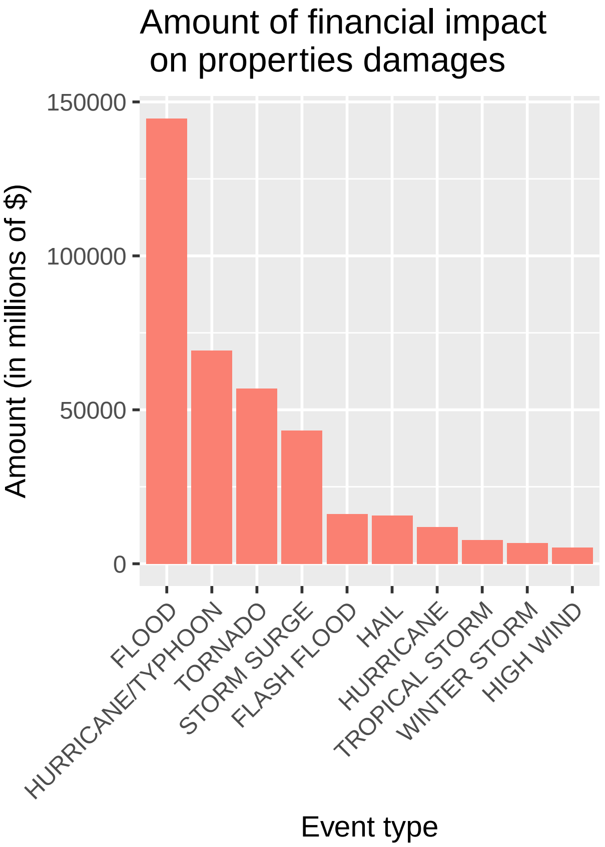


For both impacts on population (fatalities and injuries), tornado is the event type with the highest impact, some way ahead the othe event types.

Question 2 : Across the United States, which types of events have the greatest economic consequences?

Here we will draw a table giving the 10 major event types and their financial impact. This time, we will use bar plots.





The event type with the highest impact is :

- drought for crop damages,
- flood for properties damages