Storm events impact on population health and economy

Fabien Nugier

11/12/2019

Synopsis

This project is the second assigment of the **Coursera lecture "Reproducible Research"** by the **Johns Hopkins University**. In this project, we will consider data from the *U.S. National Oceanic and Atmospheric Administration (NOAA)* database in order to answer two questions:

- Across the United States, which types of events are most harmful with respect to population health?
- Across the United States, which types of events have the greatest economic consequences?

We will first consider the data and explain how it is transformed, we will then present the results and short conclusions. This project was carried out on **Tuesday Nov 12, 2019**, using RStudio with R markdowns.

Data Processing

We first load the single file directly from .csv.bz2 (to have a 50MB file instead of 500MB on the local machine):

```
d0 <- read.csv("./repdata_data_StormData.csv.bz2", header=TRUE, sep=",")
```

To understand better the data we have, let us extract some basic information necessary to answer our questions. The dimension of the table is:

```
dim(d0)
```

```
## [1] 902297     37
```

and the names of the columns are:

```
names(d0)
```

```
##  [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"
```

Events

The list of events contained in `d0$EVTYPE` is as follows:

```
eventtypes <- unique(d0$EVTYPE)
head(eventtypes,n=10)
```

```
##  [1] TORNADO                   TSTM WIND
##  [3] HAIL                      FREEZING RAIN
##  [5] SNOW                      ICE STORM/FLASH FLOOD
##  [7] SNOW/ICE                  WINTER STORM
```

```
##  [9] HURRICANE OPAL/HIGH WINDS THUNDERSTORM WINDS
## 985 Levels:    HIGH SURF ADVISORY  COASTAL FLOOD ... WND
```

and the number of them is:

```
length(eventtypes)
```

```
## [1] 985
```

Population Health

We will use the columns `d0$FATALITIES` and `d0$INJURIES` to evaluate the impact of storm events on population health. These columns contain integer values corresponding respectively to the number of people killed or injured by storm events. Here is a summary of these columns:

```
summary(d0$FATALITIES)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   0.0000   0.0000   0.0000   0.0168   0.0000 583.0000
```

```
summary(d0$INJURIES)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##    0.0000    0.0000    0.0000    0.1557    0.0000 1700.0000
```

Economy

To evaluate the impact of events on the economy, we consider the four following columns: - `d0$PROPDMG` : property damage - `d0$PROPDMGEXP` : symbom associated ('K' for kilo, 'M' for million, 'B' for billion) - `d0$CROPDMG` : crops damage - `d0$CROPDMGEXP` : symbom associated ('K' for kilo, 'M' for million, 'B' for billion) and we should add that these values are expressed in USD (US dollars).

NA values

We can see that the file contains a lot of NA values:

```
sum(is.na(d0))
```

```
## [1] 1745947
```

but none are in the columns we are interested about:

```
sum(is.na(d0[,c("FATALITIES","INJURIES","PROPDMG","PROPDMGEXP","CROPDMG","CROPDMGEXP")]))
```

```
## [1] 0
```

which will make our task easier.

Removing useless columns

In order to lighten the dataframe, we remove the columns that we dont need and define a new dataframe with only the rows that we may use after:

```
d <- d0[c("BGN_DATE","BGN_TIME","TIME_ZONE","LATITUDE","LONGITUDE","STATE","EVTYPE","MAG","FATALITIES",
rm(d0)
```

Time

The time is given by the 3 columns `d$BGN_DATE`, `d$BGN_TIME` and `d$TIME_ZONE`. It can be interesting to have the time of events in order to look for correlations in time. However, the record of timezones is quite complex and most of the values are not recognized by the function `strptime`. The values are the following:

```
summary(d$TIME_ZONE)
```

```
##     ADT     AKS     AST     CDT     CSC     CSt     CST     EDT     ESt     EST
##       3    1553    6360     692       1       4  547493     569       2  245558
##     ESY     GMT     GST     HST     MDT     MST     PDT     PST     SCT     SST
##       1       1      32    2563      99   68390     154   28302       2     505
##     UNK     UTC
##       9       4
```

Consequently, we drop the time zones from the table. And since these zones are not accounted for, there is no use accounting for the time neither. We hence keep only the dates, knowing that the time zones of record may actually put the dates off by +/- a day. We also convert the dates into date types:

```
d$BGN_DATE <- as.Date(as.character(d$BGN_DATE),format="%m/%d/%Y")
d <- d[c("BGN_DATE","LATITUDE","LONGITUDE","STATE","EVTYPE","MAG","FATALITIES","INJURIES","PROPDMG","PR
```

Renaming columns and display first rows

Finally, we rename the columns to make them easier to handle:

```
names(d) <- tolower(names(d))
names(d)[1] <- "date"
names(d)[2] <- "lat"
names(d)[3] <- "long"
names(d)[7] <- "fatal"
names(d)[8] <- "injur"
head(d)
```

```
##         date  lat long state  evtype mag fatal injur propdmg propdmgexp
## 1 1950-04-18 3040 8812    AL TORNADO   0     0    15    25.0          K
## 2 1950-04-18 3042 8755    AL TORNADO   0     0     0     2.5          K
## 3 1951-02-20 3340 8742    AL TORNADO   0     0     2    25.0          K
## 4 1951-06-08 3458 8626    AL TORNADO   0     0     2     2.5          K
## 5 1951-11-15 3412 8642    AL TORNADO   0     0     2     2.5          K
## 6 1951-11-15 3450 8748    AL TORNADO   0     0     6     2.5          K
##   cropdmg cropdmgexp
## 1       0
## 2       0
## 3       0
## 4       0
## 5       0
## 6       0
```

The data is ready !

Results

We will still have to manipulate the final dataframe and extract some plots out of it, so we load the necessary libraries for doing so:

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(lattice)
```

Population Health

Let us focus on fatalities and injuries and create a table `d1` that contains them in order of decreasing victims. We then split it into a table for fatalities and one for injuries, so that we can sort them separately:

```
d1 <- d %>% group_by(evtype) %>% summarise(TotFatal=sum(fatal),TotInjur=sum(injur))

d1Fatal <- d1[c("evtype","TotFatal")]
d1Injur <- d1[c("evtype","TotInjur")]

d1Fatal <- d1Fatal[order(d1Fatal$TotFatal, decreasing = TRUE),]
d1Injur <- d1Injur[order(d1Injur$TotInjur, decreasing = TRUE),]

d1Fatal$evtype <- factor(d1Fatal$evtype, levels=d1Fatal$evtype[order(d1Fatal$TotFatal)])
d1Injur$evtype <- factor(d1Injur$evtype, levels=d1Injur$evtype[order(d1Injur$TotInjur)])
```

We can plot this figure with `ggplot2` and `gridExtra` in order to have the:

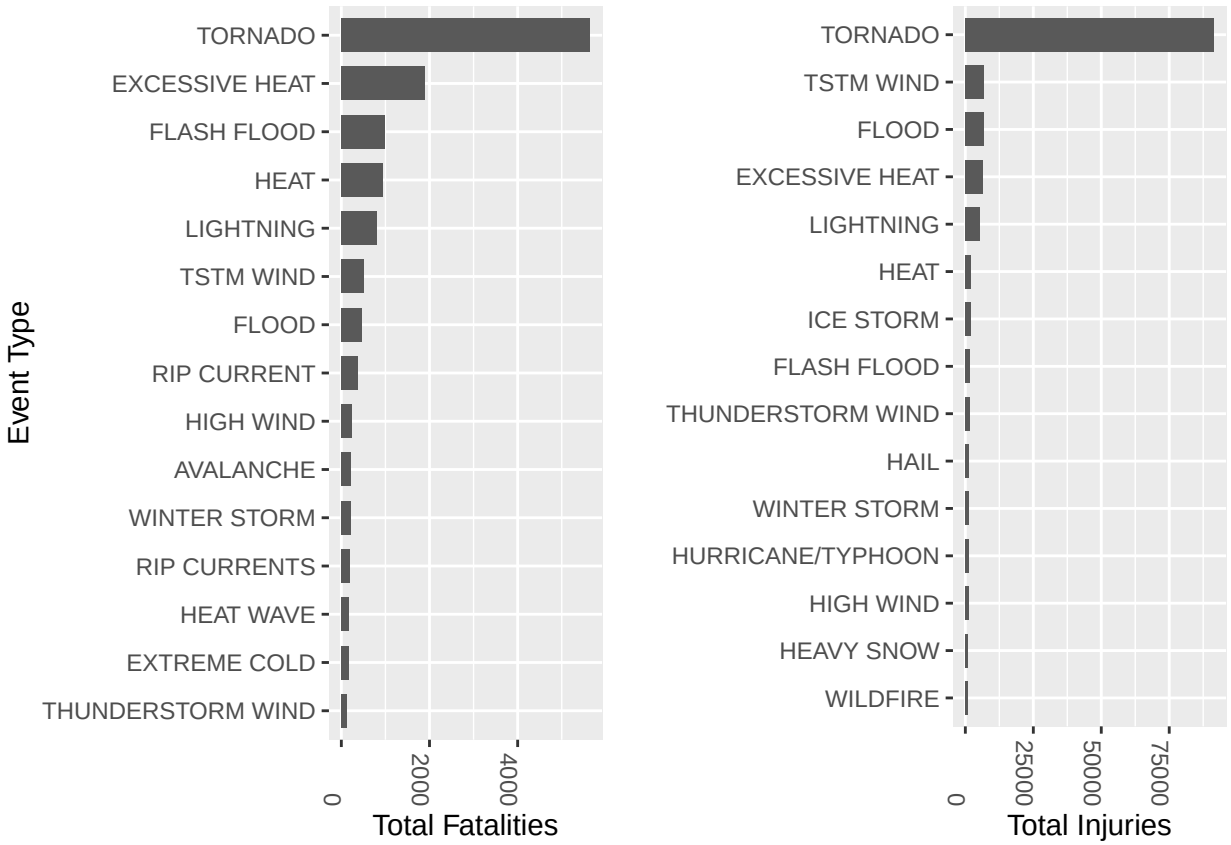
```
require(gridExtra)
```

```
## Loading required package: gridExtra
```

```
##
## Attaching package: 'gridExtra'
```

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

```
p1 <- ggplot(d1Fatal[1:15,], aes(x=evtype, y=TotFatal)) + geom_col(width=0.7) + coord_flip() + theme(ax
p2 <- ggplot(d1Injur[1:15,], aes(x=evtype, y=TotInjur)) + geom_col(width=0.7) + coord_flip() + theme(ax
grid.arrange(p1,p2,ncol=2)
```



We can see the fatalities are mainly caused by tornatos, excessive heats, flash floods, etc. Injuries are also mainly caused by tornados, but the second factor of injuries is thurderstorms winds and then floods.

This answers the question of which types of events are the most harmful to population health in the United States.

Economy

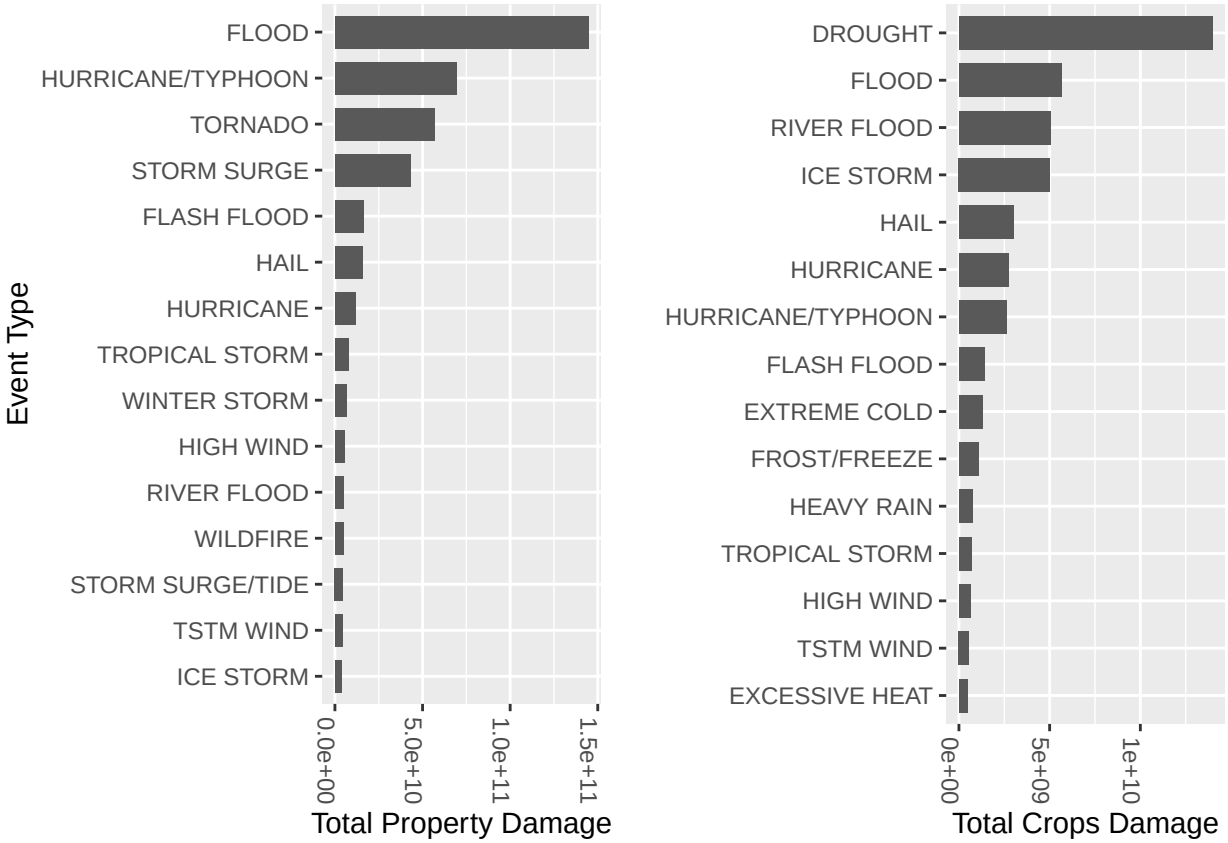
Let us build the dataframe containing the property damages and crop damages and correct the data according to the indication we get from `propdmgexp` and `propdmgexp`. We then proceed in the same way as for the previous plot, sorting the data by importance of damage:

```
d2 <- d[,c("evtype","propdmg","propdmgexp","cropdmg","cropdmgexp")]
d2$propdmg[d2$propdmgexp=="K"] <- d2$propdmg[d2$propdmgexp=="K"] * 1E3
d2$propdmg[d2$propdmgexp=="M"] <- d2$propdmg[d2$propdmgexp=="M"] * 1E6
d2$propdmg[d2$propdmgexp=="B"] <- d2$propdmg[d2$propdmgexp=="B"] * 1E9
d2$cropdmg[d2$cropdmgexp=="K"] <- d2$cropdmg[d2$cropdmgexp=="K"] * 1E3
d2$cropdmg[d2$cropdmgexp=="M"] <- d2$cropdmg[d2$cropdmgexp=="M"] * 1E6
d2$cropdmg[d2$cropdmgexp=="B"] <- d2$cropdmg[d2$cropdmgexp=="B"] * 1E9

d2 <- d2 %>% group_by(evtype) %>% summarise(TotPropDmg=sum(propdmg),TotCropDmg=sum(cropdmg))
d2Prop <- d2[c("evtype","TotPropDmg")]
d2Crop <- d2[c("evtype","TotCropDmg")]
d2Prop <- d2Prop[order(d2Prop$TotPropDmg, decreasing = TRUE),]
d2Crop <- d2Crop[order(d2Crop$TotCropDmg, decreasing = TRUE),]

d2Prop$evtype <- factor(d2Prop$evtype, levels=d2Prop$evtype[order(d2Prop$TotPropDmg)])
```

```
d2Crop$evtype <- factor(d2Crop$evtype, levels=d2Crop$evtype[order(d2Crop$TotCropDmg)])
```

We can then plot the two results along each other:

```
require(gridExtra)
p3 <- ggplot(d2Prop[1:15,], aes(x=evtype, y=TotPropDmg)) + geom_col(width=0.7) + coord_flip() + theme(a
p4 <- ggplot(d2Crop[1:15,], aes(x=evtype, y=TotCropDmg)) + geom_col(width=0.7) + coord_flip() + theme(a
grid.arrange(p3,p4,ncol=2)
```



We see that the property damage is dominated by floods, hurricanes and tornados. As for the crops damages, it is first the droughts, then floods and river floods which have the most impact.

This answers the question of which types of events have the greatest economic consequences over the United States.

Bonus: Strength of the events with time

As a bonus, let us exploit the time variable and see if there is an effect since 50 years of events getting stronger because of global warming. Note here that there may be a sample bias here if the rate of record of events is increasing with time, since we are going to add casualties and damage costs.

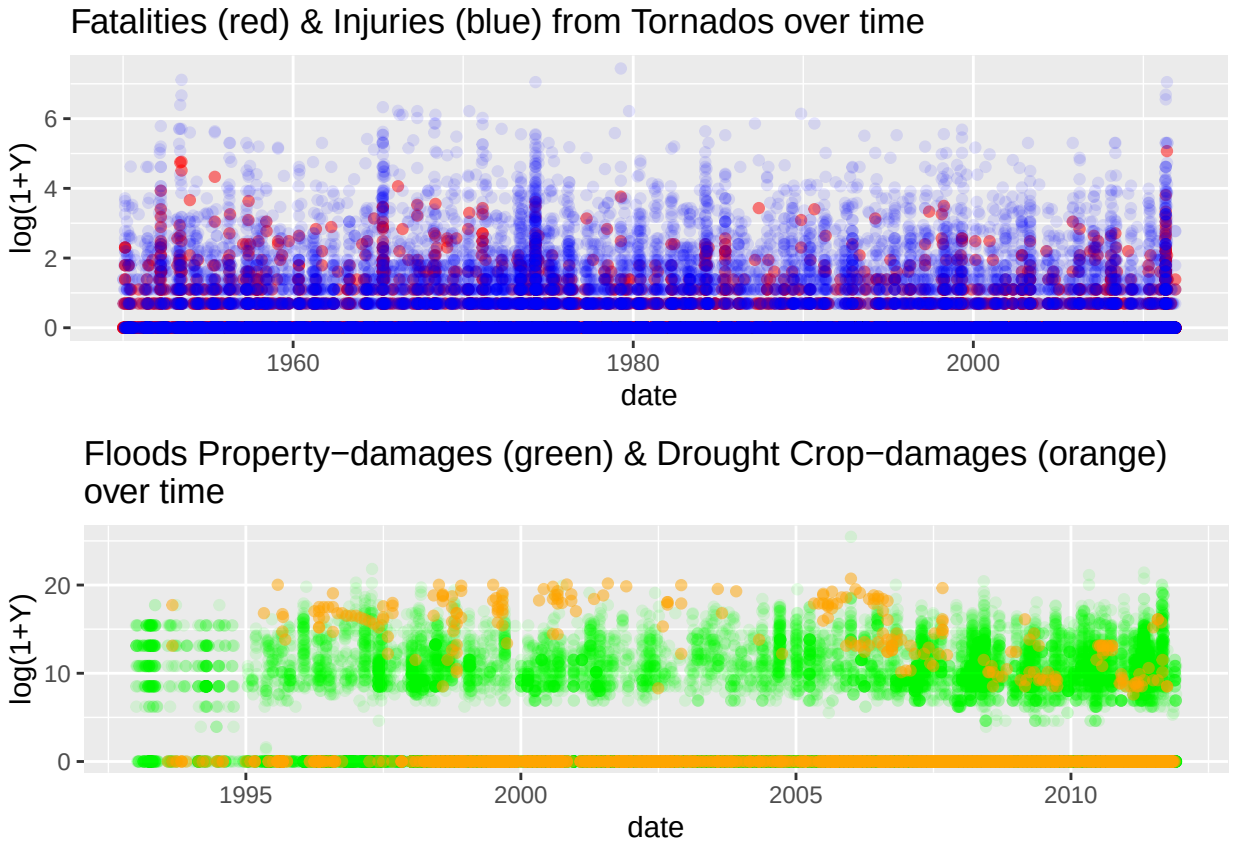
Let us build the tables:

```
# fatalities/injuries by tornado
d3 <- d[c("date","evtype","fatal","injur")]
d3 <- d3[d3$evtype=="TORNADO",]
# property damage by flood
d4 <- d[c("date","evtype","propdmg","propdmgexp")]
```

```
d4$propdmg[d4$propdmgexp=="K"] <- d4$propdmg[d4$propdmgexp=="K"] * 1E3
d4$propdmg[d4$propdmgexp=="M"] <- d4$propdmg[d4$propdmgexp=="M"] * 1E6
d4$propdmg[d4$propdmgexp=="B"] <- d4$propdmg[d4$propdmgexp=="B"] * 1E9
d4 <- d4[d4$evtype=="FLOOD",]
# crop damage by drought
d5 <- d[c("date","evtype","cropdmg","cropdmgexp")]
d5$cropdmg[d5$cropdmgexp=="K"] <- d5$cropdmg[d5$cropdmgexp=="K"] * 1E3
d5$cropdmg[d5$cropdmgexp=="M"] <- d5$cropdmg[d5$cropdmgexp=="M"] * 1E6
d5$cropdmg[d5$cropdmgexp=="B"] <- d5$cropdmg[d5$cropdmgexp=="B"] * 1E9
d5 <- d5[d5$evtype=="DROUGHT",]
```

And we get the corresponding plots combined together:

```
require(gridExtra)
p5 <- ggplot() + geom_point(data=d3, aes(x=date, y=log(1+fatal)), color="red", alpha=0.5) + geom_point(
p6 <- ggplot() + geom_point(data=d4, aes(x=date, y=log(1+propdmg)), color="green", alpha=0.1) + geom_poi
grid.arrange(p5,p6,nrow=2)
```



We cannot really see a clear increase in fatalities/injuries or damages here. This may be due to bias in the measurements or just better prevention that compensate for the losses. One large limitation here is that we are focusing our study to the most impactful storm event for each category of casualty or damage, while we should consider them all together. Since the assigment constraints required no more than 3 plots, we will let that goal for a potential later study.

Conclusions

We have considered storm events data from the U.S. National Oceanic and Atmospheric Administration (NOAA) database. We have shown that the events that caused the highest number of fatalities was tornados, followed by excessive heat and flash floods. As for injuries, it is also tornados coming first, but follwed by thunderstorm winds and floods. We then showed that floods, followed by hurricanes and tornados cause the most damages to property. On the other hand, it is droughts followed by floods and river floods which cause most damages to crops. As an extra study, we have plotted these fatalities/injuries and damages against time for the dominant storm event, but the results where not so conclusive on a time-dependent evolution, and as such require more investigation.