Tornadoes and floods : the most harmful and expensive disasters in the US Author : Joanne Breitfelder

Synopsis

The present project aims at studiyng the impact of severe weather events on public health and economy in the US. To find what events are the most harmful and have the greatest economic consequences, we will explore the U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database (More information here). This database tracks characteristics of major storms and weather events in the United States, including when and where they occur, as well as estimates of any fatalities, injuries, and property damage. We will first process the database to make it smaller and clearer (in particular, we will remove non-relevant data, and slightly transform some variables). We will then analyse the data and answer the question, firstly in terms of public health, and secondly in terms of economic consequences.

Data processing

Here are the principal steps we will follow to clean and tidy the data frame :

- Loading useful R packages
- Downloading, unzipping and reading the data
- Selecting the relevant variables
- Cleaning the crop and property damage variables
- Renaming the variables

Loading useful R packages

```
library(R.utils)
library(dplyr)
library(reshape2)
library(ggplot2)
library(knitr)
```

Downloading, unzipping and reading the data

```
url <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
if (!file.exists("repdata-data-StormData.bz2")) {
        download.file(url, destfile="repdata-data-StormData.bz2", method="curl")
        bunzip2("repdata-data-StormData.bz2", "StormData.csv", skip=TRUE)}
```

```
## [1] "StormData.csv"
## attr(,"temporary")
## [1] FALSE
```

```
## Load the data will likely take a few seconds. Be patient!
Storm_data_original <- read.csv("StormData.csv", na.strings=c("", "NA"))
```

Keeping only variables of interest

The original table contains 37 variables but only a few of them are of direct interest in the present study. After subsetting the table, we create a new variable containing the total number of victims per event (fatalities and injuries).

```
Storm_data <- select(Storm_data_original, BGN_DATE, STATE, EVTYPE, FATALITIES:CROPDMGEXP)
Storm_data <- mutate(Storm_data, all_health=FATALITIES+INJURIES,
                     CROPDMGEXP=as.character(CROPDMGEXP), PROPDMGEXP=as.character(PROPDMGEXP))
```

Creating a year variable

We will change the date format to keep only the year. This variable will be used at the end of the present study, to plot a time series of the data.

```
time_data <- colsplit(Storm_data$BGN_DATE, " ", c("date", "time"))
time_data <- colsplit(time_data$date, "/", c("month", "day", "year"))
Storm_data <- mutate(Storm_data, BGN_DATE=time_data$year)
```

Cleaning the crop and property damage variables

Crop (respectively property) damage data are given in 2 columns : the first one gives the cost generated by the event, and the second contains a symbol indicating weither the cost is given in hundreads (“h” or “H”), kilos (“k” or “K”), millions (“m” or “M”) or billions (“B”) of dollars. This variable contains other symbols (e.g. “+”, “-”, “?”,.) that we will not consider in the present study. This filtering concerns only 0.3% of the data, so we can assume that it will not introduce any significant bias in the final result.

After calculating the cost generated by crop (resp. property) damage for each event, we create a new variable containing the total cost generated by the event.

```
mult <- c("k", "K", "m", "M", "B", "h", "H", NA)
new_mult = c(1e3, 1e3, 1e6, 1e6, 1e9, 1e2, 1e2, 1)

## Keeping only valid values of PROPDMGEXP (resp. CROPDMGEXP) :
Storm_data <- filter(Storm_data, CROPDMGEXP %in% mult, PROPDMGEXP %in% mult)

## Changing the "mult" indicators into actual conversion factors :
Storm_data$CROPDMGEXP[is.na(Storm_data$CROPDMGEXP)] <- 1
Storm_data$PROPDMGEXP[is.na(Storm_data$PROPDMGEXP)] <- 1
for (i in 1:length(mult)) {
        Storm_data$CROPDMGEXP[Storm_data$CROPDMGEXP==mult[i]] <- new_mult[i]
        Storm_data$PROPDMGEXP[Storm_data$PROPDMGEXP==mult[i]] <- new_mult[i]}

## Converting all the costs in dollars, and calculating the total cost (crop+property damage)
Storm_data <- mutate(Storm_data, CROPDMG=CROPDMG*as.numeric(CROPDMGEXP),
                     PROPDMG=PROPDMG*as.numeric(PROPDMGEXP), eco=CROPDMG+PROPDMG)
Storm_data <- select(Storm_data, -CROPDMGEXP, -PROPDMGEXP)
```

Renaming the variables to get a tidy dataframe

```
names(Storm_data) <- c("year", "state", "event_type", "fatalities", "injuries",
                       "property_damage", "crop_damage", "all_health", "all_eco")
```

We have now a tidy dataset, ready to be used.

```
kable(head(Storm_data, n=10))
```

year	state	event_type	fatalities	injuries	property_damage	crop_damage	all_health	all_eco
1950	AL	TORNADO	0	15	25000	0	15	25000
1950	AL	TORNADO	0	0	2500	0	0	2500
1951	AL	TORNADO	0	2	25000	0	2	25000
1951	AL	TORNADO	0	2	2500	0	2	2500
1951	AL	TORNADO	0	2	2500	0	2	2500
1951	AL	TORNADO	0	6	2500	0	6	2500
1951	AL	TORNADO	0	1	2500	0	1	2500
1952	AL	TORNADO	0	0	2500	0	0	2500
1952	AL	TORNADO	1	14	25000	0	15	25000
1952	AL	TORNADO	0	0	25000	0	0	25000

Results

The most harmful events

Here are the different parts of the code that we will execute to extract interesting results from our tidy data frame.

- Calculating the total number of victims for each event :

```
new_table <- group_by(Storm_data, event_type)
new_table <- summarize(new_table, fatalities=sum(fatalities),
                       injuries=sum(injuries), all=sum(all_health))
```

- Finding the 10 worst events (highest numbers of victims) :

```
all_table <- arrange(select(new_table, event_type, all), desc(all))
worst_events <- head(all_table, n=10)$event_type
```

- Turning the 3 variables related to victims (“fatalities”, “injuries” and “all”) into one single 3-levels factor named “category” :

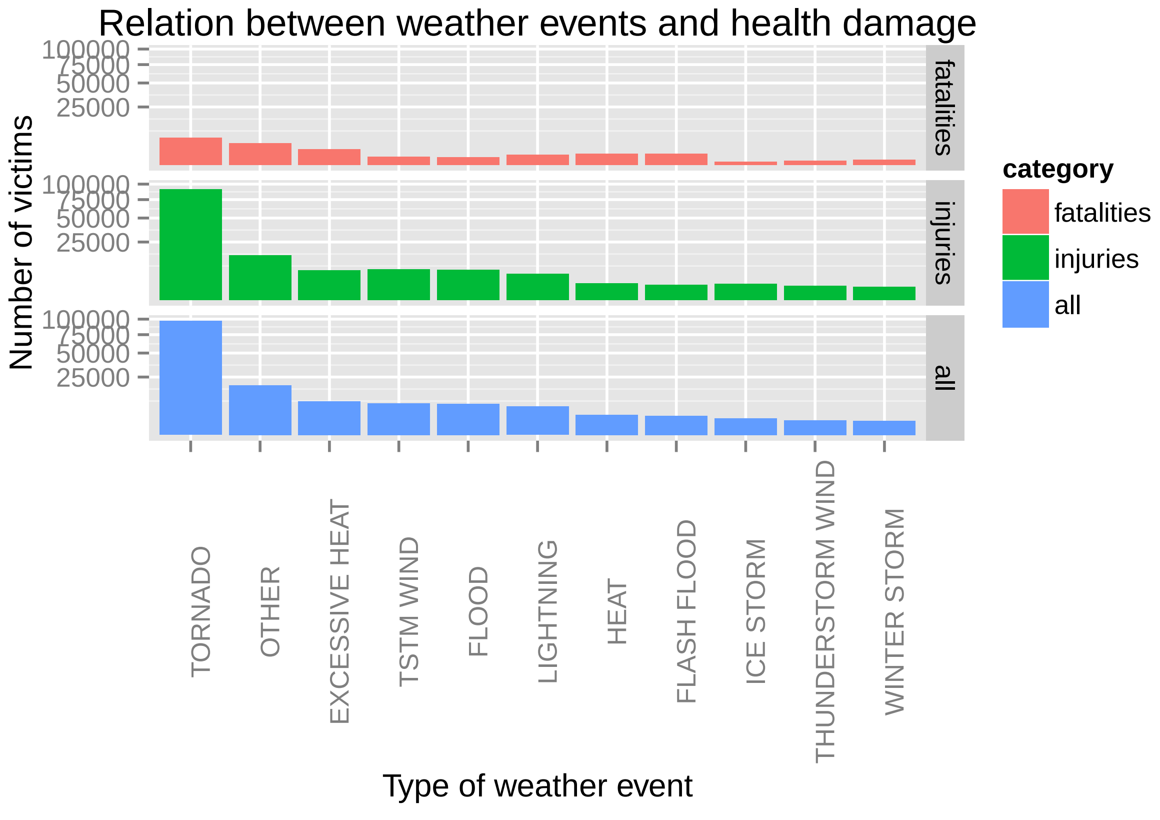
```
new_table <- melt(new_table, id.vars="event_type")
names(new_table) <- c("event_type", "category", "number")
```

- Grouping all events that are not among the 10 worst ones in a group called “OTHER” :

```
cond <- new_table$event_type %in% worst_events
new_table <- mutate(new_table,
                    event_type=ifelse(cond, as.character(event_type), "OTHER"))

# Summing all victims in each category :
new_table <- group_by(new_table, event_type, category)
new_table <- summarize(new_table, number=sum(number))
```

- Plotting the results :

```
plot <- ggplot(new_table, aes(x=reorder(event_type, -number), y=number, fill=category)) +
        geom_bar(stat="identity", position="dodge") +
        facet_grid(category~.) + theme(axis.text.x=element_text(angle=90)) +
        xlab("Type of weather event") + ylab("Number of victims") +
        ggtitle("Relation between weather events and health damage") +
        scale_y_sqrt()
print(plot)
```



From this graphic, we can see that the most harmful climatic events in the US are **tornadoes** (both in terms of injuries and fatalities), followed by **excessive heat**. For the 10 most harmful events, let's identify the states that are the most affected (e.g. for which the total number of victims in the highest) :

```
new_table <- filter(Storm_data, event_type %in% worst_events)
new_table <- group_by(new_table, state, event_type)
new_table <- summarize(new_table, all_health=sum(all_health))
new_table <- group_by(new_table, event_type)
new_table <- summarize(new_table, state=state[which.max(all_health)])
kable(new_table)
```

event_type	state
EXCESSIVE HEAT	MO
FLASH FLOOD	TX
FLOOD	TX
HEAT	IL
ICE STORM	OH
LIGHTNING	FL
THUNDERSTORM WIND	IL
TORNADO	TX
TSTM WIND	TX
WINTER STORM	UT

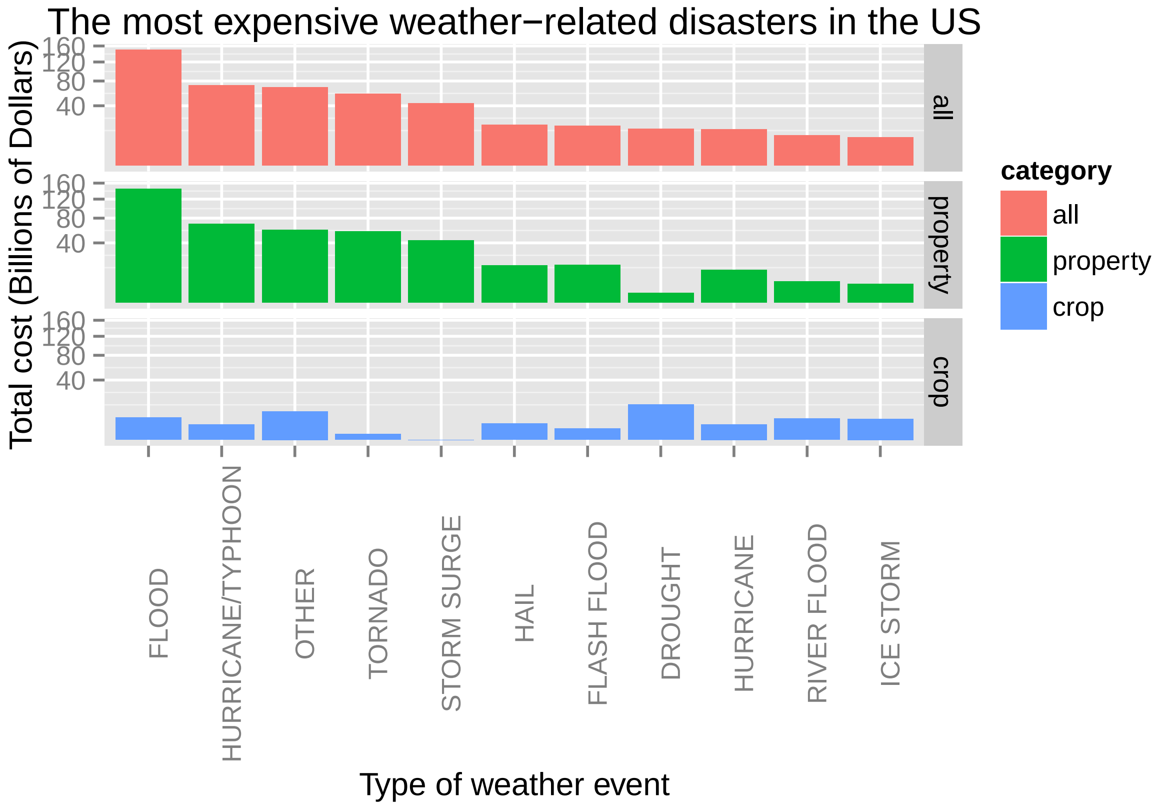
We can see that **Texas** is particularly affected by weather disasters, as it shows the highest number of victims for **flash and slow floods**, **tornadoes** and **thunderstorm winds** (i.e. 4 of the 10 most harmful disasters experienced by the whole country).

Economic consequences

The computing analysis being quite similar to the one used to study the quantity of victims, we will display the code without more explanations.

```
#-------------------------------------------------------
# Calculating the total damage for each event
new_table <- group_by(Storm_data, event_type)
new_table <- summarize(new_table, all=sum(all_eco),
                       property=sum(property_damage), crop=sum(crop_damage))

#-------------------------------------------------------
# Finding the 10 worst events : highest number of victims (injuries + fatalities)
all_table <- arrange(select(new_table, event_type, all), desc(all))
worst_events <- head(all_table, n=10)$event_type

#-------------------------------------------------------
# Turning the 3 variables related to damages into one single 3-levels factor :
new_table <- melt(new_table, id.vars="event_type")
names(new_table) <- c("event_type", "category", "number")

#-------------------------------------------------------
# Grouping all events that are not in the worst ones in a category "OTHER"
cond <- new_table$event_type %in% worst_events
new_table <- mutate(new_table,
                    event_type=ifelse(cond, as.character(event_type), "OTHER"))

#-------------------------------------------------------
# Summing all victims in the category OTHER
new_table <- group_by(new_table, event_type, category)
new_table <- summarize(new_table, number=sum(number))
new_table <- arrange(new_table, category, desc(number))
```

```
#------------------------------------------------------
# Plotting the results :)
plot <- ggplot(new_table, aes(x=reorder(event_type, -number), y=number/1e9, fill=category)) +
        geom_bar(stat="identity", position="dodge") +
        facet_grid(category~.) + theme(axis.text.x=element_text(angle=90)) +
        xlab("Type of weather event") + ylab("Total cost (Billions of Dollars)") +
        ggtitle("The most expensive weather-related disasters in the US") +
        scale_y_sqrt()
print(plot)
```



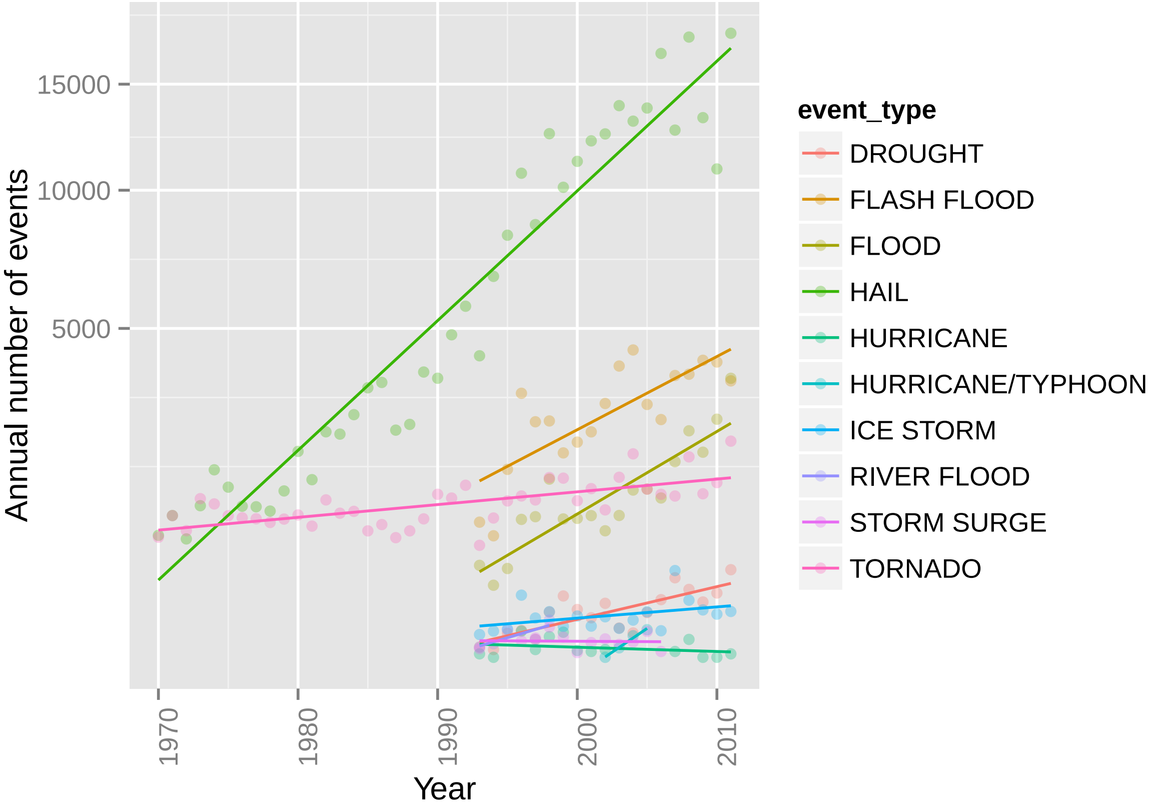
We can see that the most expensive extreme weather events in the US (integrated over the past ~60 years) are **floods** and **hurricanes** (in terms of property damage). These disasters can damage buildings and bigger structures such as bridges, roadways or electric installations. In terms of crop damage (a very important economic issue in the US), the worst events are **droughts** and **floods** (both regular and river floods). They events can indeed have a devastator impact on the cultures.

Adapting a future strategy

It is known that human activities have caused important changes in weather patterns in the last decades. How is evolving the frequency at which the 10 most expensive events occur? Let's figure that out by plotting a time series. Here we considere only the data from 1970 to 2011. Indeed, in the earlier years of the database there are generally fewer events recorded.

```
new_table <- filter(Storm_data, event_type %in% worst_events, year %in% 1970:2015)
new_table <- group_by(new_table, year, event_type)
new_table <- summarize(new_table, all=length(event_type))
```

```
plot <- ggplot(new_table, aes(x=year, y=all)) +
        geom_point(aes(color=event_type), alpha=0.3) +
        geom_smooth(aes(color=event_type), se=FALSE, fullrange=FALSE, method="lm") +
        ylab("Annual number of events") + xlab("Year") +
        theme(axis.text.x=element_text(angle=90)) +
        scale_y_sqrt()
print(plot)
```



We have to keep in mind that a small bias can exist in this plot, due to the fact that more and more events are recorded. However, the trends are clear enough to conclude : almost all the events occur more and more frequently! It is very clear in particular for **hails**, **flash floods**, **floods** and **droughts**. We therefore have to adapt an economic strategy that takes into account this tendency.