The most catastrophic weather events in the U.S.

Max Almodovar

Sunday, August 23, 2015

In this project we explore the U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database (https://d396qusza40orc.cloudfront.net/repdata%2Fpeer2_doc%2Fpd01016005curr.pdf). This database contains information related to some of the major weather events in the United States occured from 1950 to 2011. The dataset includes specific information such as the type of event, when the event occured, as well as estimates of any fatalities, injuries, and property damage.

The objective of this analysis is to determine which events are the most catastrophic events in the U.S. in terms of property damage, fatalities, and injuries. First section explains how the original dataset was prepared for analytical purposes. After that, we describe the most catastrophic events analyzing the property/human life costs, as well as the evolution over time of these events.

The following library will be used for this project:

```
library(dplyr)
library(ggplot2)
library(knitr)
library(gridExtra)
```

1.- Data Processing

The dataset can be downloaded from the following url (https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2). The first step will be to download the dataset:

```
url <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
file <- "data.csv"

if(!file.exists(file))
        download.file(url, file, method = "curl")

raw_data <- read.csv(file, stringsAsFactors = FALSE)
```

The `raw_data` includes many variables that are not needed for this analysis. In particular we will subset the following 9 variables: REFNUM (weather event ID), BGN_DATE (when the event started), EVTYPE (event type), FATALITIES (# of fatalities), INJURIES (# of injuries), PROPDMG (estimation of property damage), PROPDMGEXP (exponential), CROPDMG (estimation of crop damage), CROPDMGEXP (exponential). In addition, we are going to apply a couple of transformations to the variables.

```
clean_data <- raw_data[,c(37, 2, 8, 23:28)]

clean_data$BGN_DATE <- as.Date(clean_data$BGN_DATE, format = "%m/%d/%Y")
clean_data$EVTYPE <- toupper(clean_data$EVTYPE)
```

2.- Exploratory Data Analysis

Now we need to analyze a little bit the clean data set. We have 902297 observations of events. However, do we need the whole dataset? The answer is no.

Since we need to analyze the most catastrophic events, we are not interested in those events that do not have an impact in properties, crops, and people. Let’s create 4 datasets:

```
property_dmg <- subset(clean_data, PROPDMG != 0)
property_dmg <- property_dmg[, c(1:3,6:7)]
crop_dmg <- subset(clean_data, CROPDMG != 0)
crop_dmg <- crop_dmg[, c(1:3,8:9)]
fatalities <- subset(clean_data, FATALITIES != 0)
fatalities <- fatalities[,  c(1:4)]
injuries <- subset(clean_data, INJURIES != 0)
injuries <- injuries[,  c(1:3, 5)]
```

What each of these datasets represent?

- `property_dmg` : represents the 239174 events that caused damage to properties.
- `crop_dmg` : represents the 22099 events that caused damage to crops.
- `fatalities` : represents the 6974 events that killed people.
- `injuries` : represents the 17604 events that injured.

2.1.- Damage caused by an event

The datasets for the property damages and the crops damages have an exponential variable. This variable is used to determine the total amount of damage in terms of hundreds (H), thousands (K), millions (M), or billions (B). However, the exponential variable contains many other values apart from these 4 letters.

After analyzing the values of the exponential variables, we decide to use only numbers so the damage will be multiplied by 10 to the power included in this exponential variable. For the values that are not numbers or letters, we will turn them into thousands.

```
property_dmg$PROPDMGEXP[property_dmg$PROPDMGEXP %in% c("K", "k", "0", "+", "-", "")] <-
"3"
property_dmg$PROPDMGEXP[property_dmg$PROPDMGEXP %in% c("m", "M")] <- "6"
property_dmg$PROPDMGEXP[property_dmg$PROPDMGEXP %in% c("h", "H")] <- "2"
property_dmg$PROPDMGEXP[property_dmg$PROPDMGEXP %in% c("B")] <- "9"
property_dmg$PROPDMGEXP <- as.numeric(property_dmg$PROPDMGEXP)
property_dmg$PROPDMGTOTAL <- property_dmg$PROPDMG * 10^property_dmg$PROPDMGEXP
property_dmg <- property_dmg[ ,-c(4,5)]

crop_dmg$CROPDMGEXP[crop_dmg$CROPDMGEXP %in% c("K", "k", "0", "")] <- "3"
crop_dmg$CROPDMGEXP[crop_dmg$CROPDMGEXP %in% c("m", "M")] <- "6"
crop_dmg$CROPDMGEXP[crop_dmg$CROPDMGEXP %in% c("B")] <- "9"
crop_dmg$CROPDMGEXP <- as.numeric(crop_dmg$CROPDMGEXP)
crop_dmg$CROPDMGTOTAL <- crop_dmg$CROPDMG * 10^crop_dmg$CROPDMGEXP
crop_dmg <- crop_dmg[ ,-c(4,5)]
```

With this code we create a variable for each dataset including the total amount of damage in USD caused by each event.

2.1.- Description of the event

From the dataset documentation (https://d396qusza40orc.cloudfront.net/repdata%2Fpeer2_doc%2Fpd01016005curr.pdf), we know that only 48 types of events are allowed. However, there are more than 300 different event types included in our 4 datasets.

First, a deep analysis of the dataset shows that there are 9 event types that cannot be classified within the standard 48 types. Since this is a very small amount, we will clean the occurences of these events.

```
invalid_events <- c("?","MARINE ACCIDENT", "MARINE MISHAP", "APACHE COUNTY",
                    "DROWNING", "HIGH", "OTHER", "URBAN AND SMALL",
                    "URBAN SMALL")
property_dmg <- subset(property_dmg, ! (EVTYPE %in% invalid_events))
crop_dmg <- subset(crop_dmg, ! (EVTYPE %in% invalid_events))
fatalities <- subset(fatalities, ! (EVTYPE %in% invalid_events))
injuries <- subset(injuries, ! (EVTYPE %in% invalid_events))
```

For the remaining events, we have a created a function named `cleaning_events()` that assigns one of the 48 standard types described in the official documentation to each of the non-standard events included in the dataset. For details on how this function works, please go to the following link (https://github.com/maxal1986/RepData_PeerAssessment2/blob/master/cleaning_events.R).

```
source("cleaning_events.R")
property_dmg$EVTYPE <- cleaning_events(property_dmg$EVTYPE)
crop_dmg$EVTYPE <- cleaning_events(crop_dmg$EVTYPE)
fatalities$EVTYPE <- cleaning_events(fatalities$EVTYPE)
injuries$EVTYPE <- cleaning_events(injuries$EVTYPE)
```

3.- Results

With the 4 datasets before we are going to apply some transformations that will be useful for reporting the information:

```
prop_events <- property_dmg %>% group_by(EVTYPE) %>% summarise(total_dmg = sum(PROPDMGTOT
AL)) %>%
        arrange(desc(total_dmg))
crop_events <- crop_dmg %>% group_by(EVTYPE) %>% summarise(total_dmg = sum(CROPDMGTOTAL))
%>%
        arrange(desc(total_dmg))
fat_events <- fatalities %>% group_by(EVTYPE) %>% summarise(killed = sum(FATALITIES)) %>%
        arrange(desc(killed))
inj_events <- injuries %>% group_by(EVTYPE) %>% summarise(injured = sum(INJURIES)) %>%
        arrange(desc(injured))
```

3.1.- Properties and Crop Damages

Top 10 most catastrophic events in terms of damages to properties

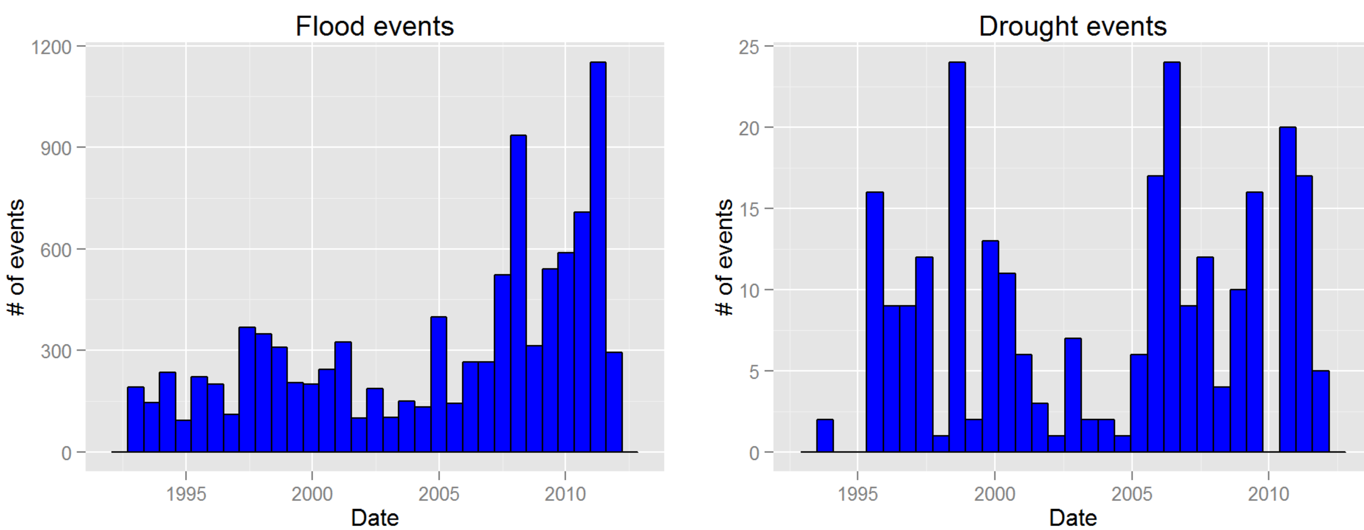
Event	Total damage in USD
FLOOD	144901469100
HURRICANE (TYPHOON)	85356410010
TORNADO	56952457780
STORM SURGE/TIDE	48036354200
FLASH FLOOD	17650829220
HAIL	17623307920
THUNDERSTORM WIND	11146232360
WILDFIRE	8496628500
TROPICAL STORM	7716127550
WINTER STORM	6689788250

Top 10 most catastrophic events in terms of damages to crops

Event	Total damage in USD
DROUGHT	13972621780
FLOOD	5672163950
HURRICANE (TYPHOON)	5516117800

LAKESHORE FLOOD	5057484000
ICE STORM	5027113500
HAIL	3114235850
FROST/FREEZE	1997061800
FLASH FLOOD	1540690250
EXTREME COLD/WIND CHILL	1335023000
THUNDERSTORM WIND	1206971650

Floods are the event with most damages to properties while drought is the event with most damages to crops. Let’s see how many events occured:



3.2.- Fatalities and Injuries

Top 10 most deadly events

Event	Total people killed
TORNADO	5633
EXCESSIVE HEAT	3132
FLASH FLOOD	1064
LIGHTNING	817
THUNDERSTORM WIND	740
RIP CURRENT	572
FLOOD	477

HIGH WIND	325
EXTREME COLD/WIND CHILL	313
AVALANCHE	269

Top 10 most people injured events

Event	**Total people injured**
TORNADO	91364
THUNDERSTORM WIND	9469
EXCESSIVE HEAT	9209
FLOOD	6791
LIGHTNING	5232
ICE STORM	2137
FLASH FLOOD	1881
HIGH WIND	1615
WILDFIRE	1608
HAIL	1467

Tornadoes are the most deadly event. It is also the event that caused more injuries.