U.S. Storm Data Analysis for Reproducible Research Assessment 2

Wien Wong

Saturday, February 17, 2015

Synopsis

This is a report created based on the analysis on U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database. The dataset includes types of the natural events, reported with respect to injuries, fatalities, economical costs of properties and crops. Besides, the latitude and longitude of the hazard events were also recorded. This report aims to analyze and find out the most serious natural disasters and its influence in U.S.

Data Processing

Grasp Rough Data

The data come in the form of a comma-separated-value file compressed via the bzip2 algorithm to reduce its size. First, import useful packages, unzip the file and read the data based on specific data frame columns.

```
echo = TRUE
library(data.table);library(knitr);library(ggplot2);library(ggmap);library(dplyr)
```

```
## Warning: package 'data.table' was built under R version 3.1.2

## Warning: package 'knitr' was built under R version 3.1.2

## Warning: package 'ggplot2' was built under R version 3.1.2

## Warning: package 'ggmap' was built under R version 3.1.2

## Warning: package 'dplyr' was built under R version 3.1.2
```

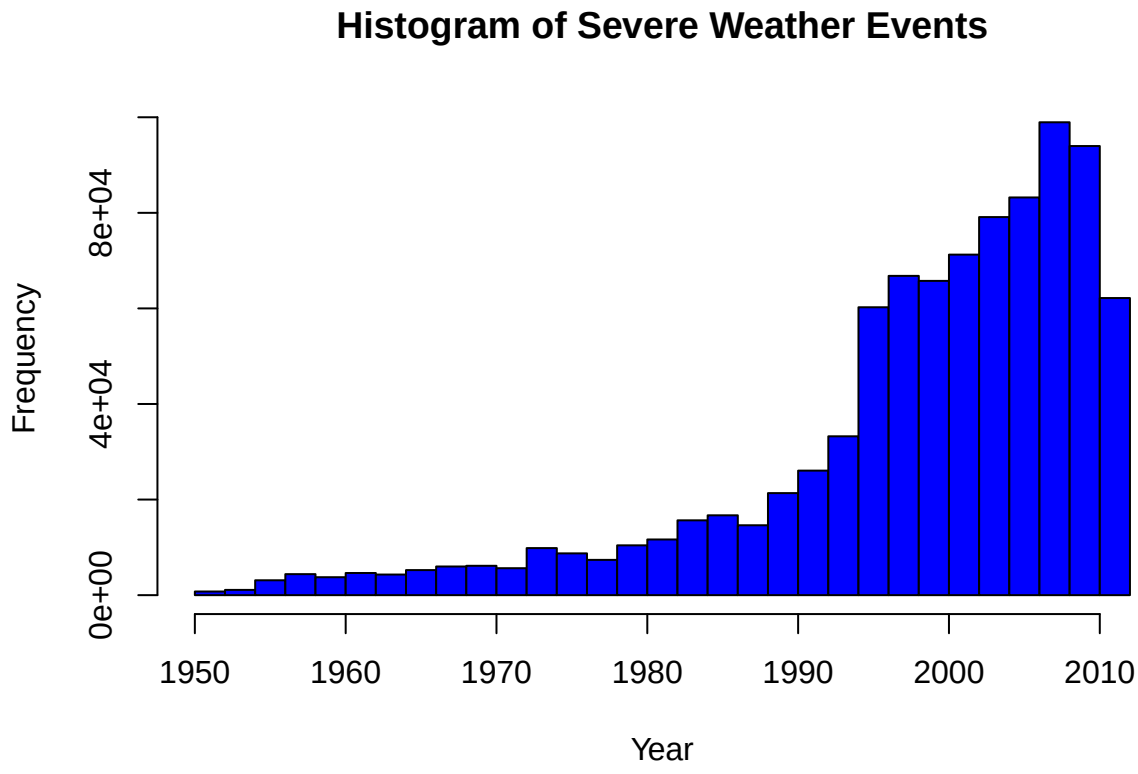
```
opts_chunk$set(cache=TRUE)
# download file
if (!file.exists("D:/Coursera_R/storm.csv.bz2")) {
    download.file("https://d396qusza40orc.cloudfront.net/
                  repdata%2Fdata%2FStormData.csv.bz2", "D:/Coursera_R/storm.csv.bz2")
}
# unzip file
if (!file.exists("D:/Coursera_R/storm.csv")) {
    library(R.utils)
    bunzip2("D:/Coursera_R/storm.csv.bz2", "D:/Coursera_R/storm.csv",
            overwrite=TRUE, remove = FALSE)
}
# read the data
storm <- data.table(read.table("storm.csv", header=T, sep=",", nrows=902298, na.strings="",
```

```
                                colClasses = c( "NULL", "character", rep("NULL", 5),
                                                "factor", rep("NULL", 14),
                                                rep("numeric", 3), "character",
                                                "numeric", "character", rep("NULL", 3),
                                                rep("numeric", 4), rep("NULL", 2) ) ))
dim(storm)
```

```
## [1] 902297     12
```

Second, let's take a look at rough tendency of the number of bad weather events from 1950 to 2011.

```
Date <- as.numeric(format(as.Date(storm$BGN_DATE, format="%m/%d/%Y %H:%M:%S"), "%Y"))
hist(Date, breaks = 30, xlab="Year", ylab="Frequency",
     main="Histogram of Severe Weather Events", col="blue")
```



As can be seen that the frequency of disasters are increased in general.

Data Subset and Clean Up

Subset the data to remove events without damage as well as discard the data with missing value.

```
storm.dmg <- filter(storm, PROPDMG !=0, CROPDMG != 0)
storm.dmg$PROPDMGEXP <- as.character(storm.dmg$PROPDMGEXP)
storm.dmg$CROPDMGEXP <- as.character(storm.dmg$CROPDMGEXP)
storm.dmg <- subset(storm.dmg, PROPDMGEXP!="" & CROPDMGEXP!="")
```

Clean up a few major event types since the data are inconsistently named.

```
storm.dmg$EVTYPE <- gsub("^.*((?!FLASH).).*FLOOD.*$","FLOOD",storm.dmg$EVTYPE,perl=TRUE)
storm.dmg$EVTYPE <- gsub("^URBAN.*$","FLOOD",storm.dmg$EVTYPE)
storm.dmg$EVTYPE <- gsub("^RIVER.*$","FLOOD",storm.dmg$EVTYPE)
storm.dmg$EVTYPE <- gsub(".*TORNADO.*","TORNADO",storm.dmg$EVTYPE)
storm.dmg$EVTYPE <- gsub("^.*THUNDER.*WIND.*$","THUNDERSTORM WIND", storm.dmg$EVTYPE)
storm.dmg$EVTYPE <- gsub(".*TSTM.*","THUNDERSTORM WIND", storm.dmg$EVTYPE)
```

Prepare Damage Data of Property/Crop

Damage data of the property and crop were converted to comparable numeric format based on meaningful units described here [1]. Both the property damage exponent and the crop damage exponent data were explored and sorted, then total damage value were calculated, separately.

```
dt <- storm.dmg
unique(dt$PROPDMGEXP)
```

```
## [1] "B" "M" "m" "K" "5" "0" "3"
```

```
dt$PROPEXP[dt$PROPDMGEXP == "K"] <- 1000
dt$PROPEXP[dt$PROPDMGEXP == "M"] <- 1e+06
dt$PROPEXP[dt$PROPDMGEXP == "B"] <- 1e+09
dt$PROPEXP[dt$PROPDMGEXP == "m"] <- 1e+06
dt$PROPEXP[dt$PROPDMGEXP == "5"] <- 1e+05
dt$PROPEXP[dt$PROPDMGEXP == "3"] <- 1000
dt$PROPEXP[dt$PROPDMGEXP == "0"] <- 1
dt$PROPDMGVAL <- dt$PROPDMG * dt$PROPEXP
#
unique(dt$CROPDMGEXP)
```

```
## [1] "M" "K" "m" "k" "B" "0"
```

```
dt$CROPEXP[dt$CROPDMGEXP == "M"] <- 1e+06
dt$CROPEXP[dt$CROPDMGEXP == "K"] <- 1000
dt$CROPEXP[dt$CROPDMGEXP == "m"] <- 1e+06
dt$CROPEXP[dt$CROPDMGEXP == "B"] <- 1e+09
dt$CROPEXP[dt$CROPDMGEXP == "k"] <- 1000
dt$CROPEXP[dt$CROPDMGEXP == "0"] <- 1
dt$CROPDMGVAL <- dt$CROPDMG * dt$CROPEXP
```

Aggregate Data by Event Type

This step aims to aggregate the data by event type w.r.t. 'FATALITIES', 'INJURIES', 'PROPDMGVAL' and 'CROPDMGVAL'.

```
fatal <- aggregate(FATALITIES ~ EVTYPE, dt, sum)
injury <- aggregate(INJURIES ~ EVTYPE, dt, sum)
casualties <- merge(fatal, injury)
propdmg <- aggregate(PROPDMGVAL ~ EVTYPE, dt, sum)
```

```
cropdmg <- aggregate(CROPDMGVAL ~ EVTYPE, dt, sum)
# top 10 event types with highest casualties
casualties10 <- casualties[order(-casualties$FATALITIES, -casualties$INJURIES), ][1:10, ]
# top 10 event types with highest property damage value
propdmg10 <- propdmg[order(-propdmg$PROPDMGVAL), ][1:10, ]
# top 10 event types with highest crop damage value
cropdmg10 <- cropdmg[order(-cropdmg$CROPDMGVAL), ][1:10, ]
```

Results

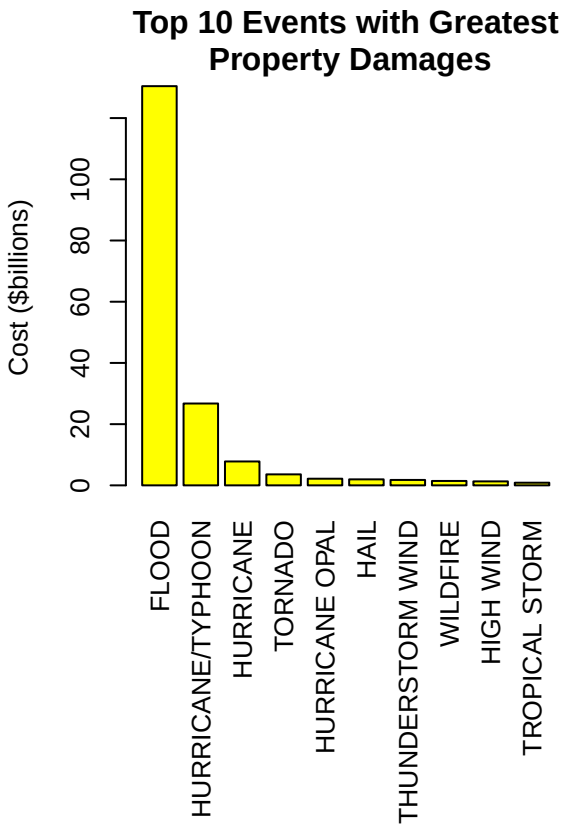
The table shown below shows the types of events impose most harmfulness w.r.t. the population health across the United States.

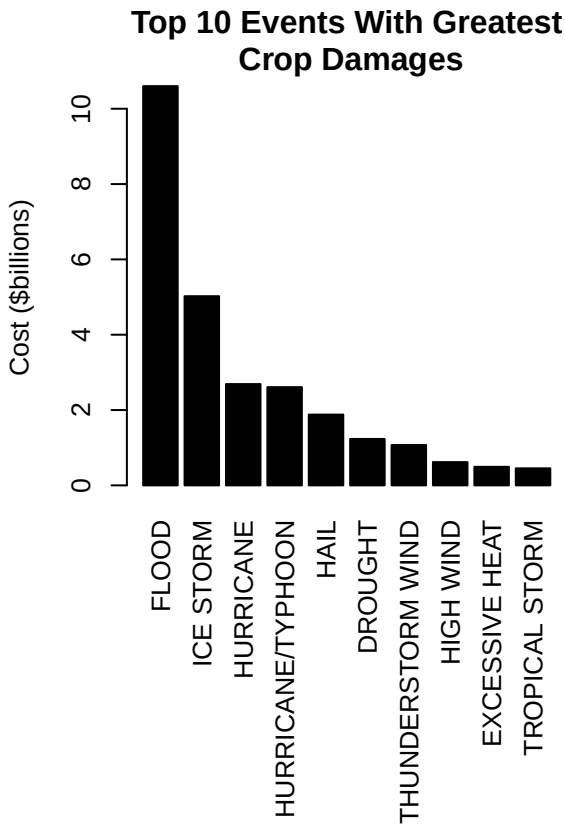
```
kable(casualties10, format = "markdown")
```

	EVTYPE	FATALITIES	INJURIES
8	FLOOD	229	6699
53	TORNADO	219	2287
6	EXCESSIVE HEAT	46	18
38	HURRICANE/TYPHOON	40	909
58	TSUNAMI	32	129
62	WILDFIRE	31	137
33	HURRICANE	29	22
22	HEAT	22	470
29	HIGH WIND	18	213
51	THUNDERSTORM WIND	16	361

The chart plotted below indicates types of events have the greatest economic consequences across the United States.

```
par(mfrow = c(1, 2), mar = c(12, 4, 3, 2), mgp = c(3, 1, 0), cex = 0.8)
barplot(propdmg10$PROPDMGVAL/(10^9), las = 3, names.arg = propdmg10$EVTYPE,
        main = "Top 10 Events with Greatest\n Property Damages",
        ylab = "Cost ($billions)",col = "yellow")
barplot(cropdmg10$CROPDMGVAL/(10^9), las = 3, names.arg = cropdmg10$EVTYPE,
        main = "Top 10 Events With Greatest\n Crop Damages",
        ylab = "Cost ($billions)", col = "black")
```

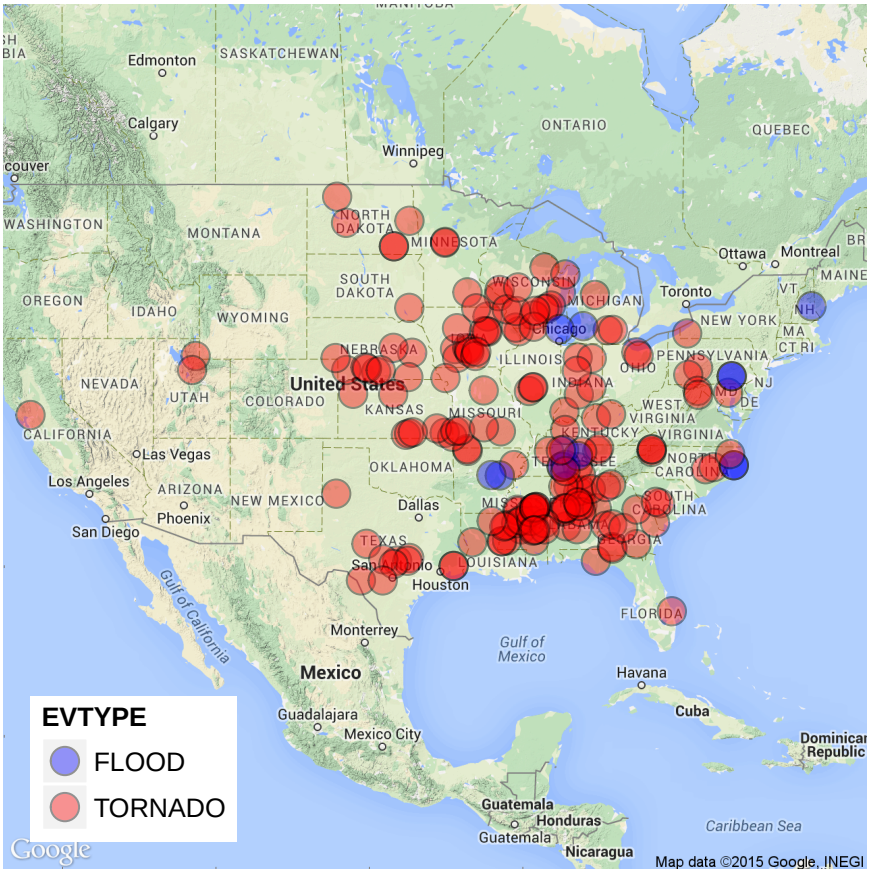




Finally, let's take a look at the spatial analysis of the harzard weather events. Here only the spatial distributions of "Flood" and "TORNADO" were examined. As shown in the figure, there are clearly more recorded tornado events than floods. Both types of events occur with high frequencies in the Central/East regions.

```
# subset with event types and damages
Selected.Events <- filter(storm, EVTYPE=="TORNADO"|EVTYPE=="FLOOD")
Selected.Events <- filter(Selected.Events, PROPDMG !=0,
                          CROPDMG != 0, FATALITIES + INJURIES !=0)
# geocode conversion: x=-LONGITUDE/100, y=LATITUDE/100
ggmap(get_map(location = 'United States', zoom = 4), extent="device") +
geom_point(data=Selected.Events, aes(x=-LONGITUDE/100, y=LATITUDE/100, fill=EVTYPE),
           color="black", pch=21, size=5, alpha=0.4) +
scale_fill_manual(values=c("blue", "red")) +
theme(legend.justification=c(0,0), legend.position=c(0,0))
```

```
## Warning: Removed 77 rows containing missing values (geom_point).
```



Conclusion

From the data, flood and tornado are most harmful w.r.t. population health, while flood, hurricane/typhoon and ice storm have the greatest economic consequences.

Reference

The cited [1] comes from this URL: http://ire.org/media/uploads/files/datalibrary/samplefiles/Storm%20Events/layout08.doc