United States Storm Impact Analysis on Economics and Health

Samuel B Baco

03/04/2020

0 Synopsis

Course Project 2 for Coursera's Reproducible Research Specialization by Johns Hopkins University.

Given the provided dataset of storm data in the United States, the student was tasked at utilizing the course materials and lessons to identify:

1. which event(s) had the greatest impact Population **health** ;
2. which event(s) had the greatest impact Population **economics**.

To do this, one must load the data, subset it, sort through injuries and fatalities (health), and then plot, as well as similar analysis for the economics.

The Economics proved more challenging, as one had to convert alphabetic variables to numeric for analysis. This lesson also taught how to publish to RPubs directly from RStudio.

1 Code preparation

To analyze the data, packages from **tidyverse** were used. In addition, **lubridade** was loaded to help with date and time variables and **grid** plus **gridExtra** were used to help plot graphics combined.

```
library(tidyverse)
library(lubridate)
library(gridExtra)
library(grid)
library(knitr)
```

2 Dowload and load data

The data for this data analysis come in the form of a comma-separated-value file compressed via the bzip2 algorithm to reduce its size.

```
destfile <- "Dataset/tmp.bz2"

if (!file.exists(destfile)) {

        fileLink <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"

        download.file(fileLink,
                      dest = destifile,
                      method = "curl")
}

weather <- read_csv("Dataset/tmp.bz2",
                    col_types = cols(BGN_DATE = col_datetime(format = "%m/%d/%Y %H:%M:%S"),
                                     BGN_TIME = col_time(format = "%H%M"),
                                     CROPDMGEXP = col_character()))
```

3 Data preparation

The original dataset has a lot of information that will not be used for this analysis. So, after load the data, only these columns were selected:

- **BGN_DATE**: the date of the event
- **EVTYPE**: type of the event
- **FATALITIES**: number of fatalities (deaths) caused by the event
- **INJURIES**: number of injuries (no death) caused by the event
- **PROPDMG** and **PROPDMGEXP**: property damage. The second variable contains the exponent for the value at first variable
- **CROPDMG** and **CROPDMGEXP**: crop damage. The second variable contains the exponent for the value at first variable

```
weatherClean <- weather %>%
        select(BGN_DATE, BGN_TIME, EVTYPE, FATALITIES, INJURIES, PROPDMG, PROPDMGEXP, CROPDMG, CROPDMGE

# Rows and columns to analyse fatalities

fatalities <- weatherClean %>%
        select(BGN_DATE, EVTYPE, FATALITIES) %>%
        filter(FATALITIES > 0)

# Rows and columns to analyse injuries

injuries <- weatherClean %>%
        select(BGN_DATE, EVTYPE, INJURIES) %>%
        filter(INJURIES > 0)

# Columns to analyse economic damage

economic <- weatherClean %>%
        select(BGN_DATE, EVTYPE, PROPDMG, PROPDMGEXP, CROPDMG, CROPDMGEXP)
```

For **economic** data, it was necessary to transform the variables PROPDMGEXP and CROPDMGEXP, from character values to numeric values, where k represented thousands, m, millions and b, billions.

```
economic <- economic %>%
        filter(grepl("m|k|b", PROPDMGEXP, ignore.case = TRUE)) %>%
        filter(grepl("m|k|b", CROPDMGEXP, ignore.case = TRUE))
```

```
economic$PROPDMGEXP <- str_replace_all(economic$PROPDMGEXP,
                                       regex(c("k" = "1e+03", "m" = "1e+06", "b" = "1e+09"),
                                             ignore_case = TRUE))

economic$CROPDMGEXP <- str_replace_all(economic$CROPDMGEXP,
                                       regex(c("k" = "1e+03", "m" = "1e+06", "b" = "1e+09"),
                                             ignore_case = TRUE))

economic <- economic %>%
        mutate(PROPDMGEXP = as.numeric(PROPDMGEXP),
                CROPDMGEXP = as.numeric(CROPDMGEXP))
```

4 Analysing data from health issues

4.1 Fatalities

To simplify the analysis, first the data from **fatalities** was filtered to TOP 5.

```
fatTop <- fatalities %>%
        group_by(EVTYPE) %>%
        summarise(fat.total = sum(FATALITIES)) %>%
        top_n(n = 5, wt = fat.total)
```

As the aggregation may create an unfair view (because not all event type may have data recorded from the same time spam), also a time series from all event types was done.

```
fatTime <- fatalities %>%
        group_by(EVTYPE, year(BGN_DATE)) %>%
        summarise(fat.total = sum(FATALITIES)) %>%
        filter(EVTYPE %in% fatTop$EVTYPE) %>%
        rename(fat.year = `year(BGN_DATE)`)
```

To compare TOP 5 event types related to fatalities, a side by side graphic was constructed. First both bar chart and time series chart were created using *ggplot2* package.

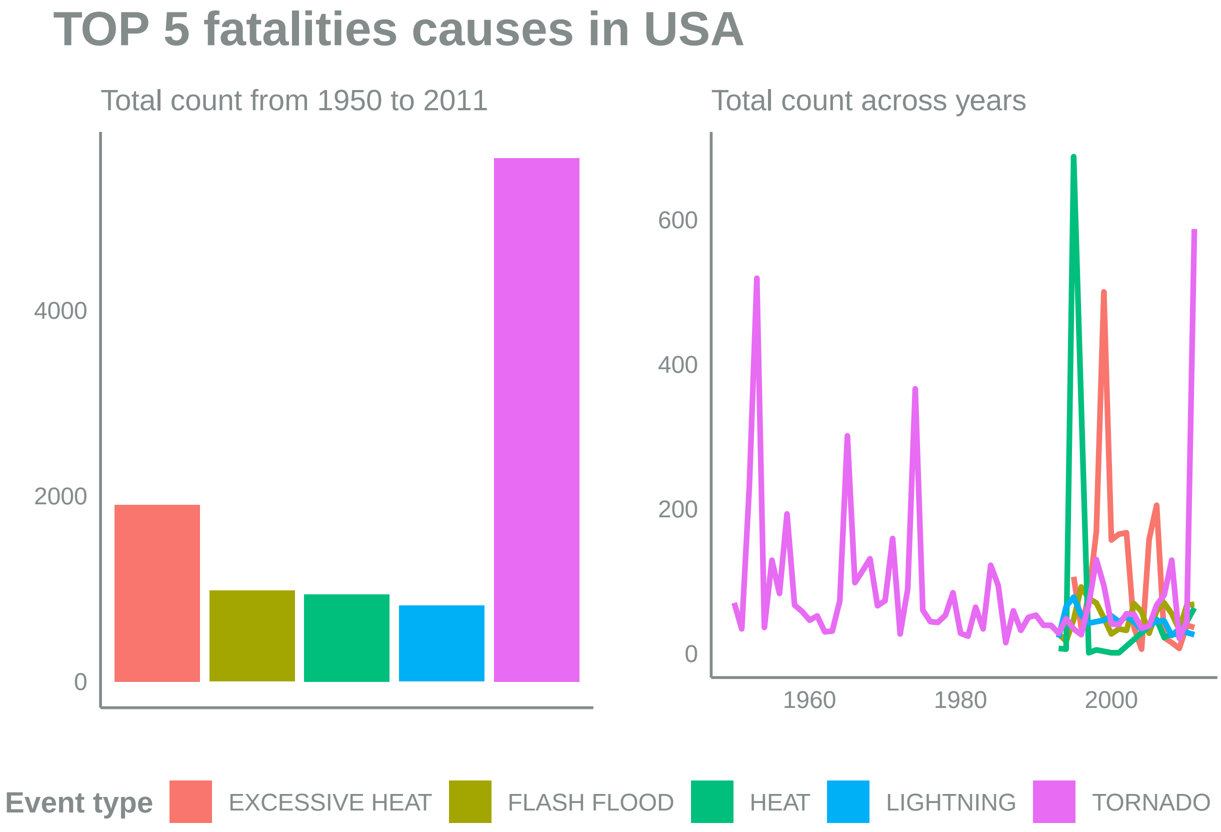
```
# TOP 5 Bar chart

fatTopPlot <- ggplot(fatTop, aes(x = EVTYPE, y = fat.total, fill = EVTYPE)) +
                geom_bar(stat = "identity") +
                        labs(title = "",
                             subtitle = "Total count from 1950 to 2011",
                             y = "",
                             x = "") +
                        scale_fill_discrete(name = "Event type") +
                        theme_minimal() +
                        theme(panel.border = element_blank(),
                              panel.grid.major = element_blank(),
                              panel.grid.minor = element_blank(),
                              axis.line = element_line(colour = "azure4"),
                              axis.title = element_text(colour = "azure4"),
                              axis.text = element_text(colour = "azure4"),
                              axis.text.x = element_blank(),
                              plot.title = element_text(colour = "azure4"),
                              plot.subtitle = element_text(colour = "azure4"),
                              strip.text = element_text(colour = "azure4"),
                              legend.text = element_text(colour = "azure4"),
                              legend.title = element_text(colour = "azure4", face = "bold"),
                              legend.position = "bottom")

# TOP 5 Time Series Chart

fYearSeries <- ggplot(fatTime, aes(x = fat.year, y = fat.total)) +
                geom_line(aes(color = EVTYPE), size = 1) +
                        labs(title = "",
                             subtitle = "Total count across years",
                             y = "",
                             x = "") +
                scale_color_discrete(name = "Event Type") +
                        theme_minimal() +
                        theme(panel.border = element_blank(),
                              panel.grid.major = element_blank(),
                              panel.grid.minor = element_blank(),
                              axis.line = element_line(colour = "azure4"),
                              axis.title = element_text(colour = "azure4"),
                              axis.text = element_text(colour = "azure4"),
                              plot.title = element_text(colour = "azure4"),
                              plot.subtitle = element_text(colour = "azure4"),
                              legend.text = element_text(colour = "azure4"),
                              legend.title = element_text(colour = "azure4", face = "bold"),
                              legend.position = "bottom")
```

A function *g_legend* was written to allow both graphics share the same legend.

```
g_legend <- function(a.gplot) {

        tmp <- ggplot_gtable(ggplot_build(a.gplot))

        leg <- which(sapply(tmp$grobs, function(x) x$name) == "guide-box")

        legend <- tmp$grobs[[leg]]

        return(legend)
}

# Get fatTopPlot legend

fatLegend <- g_legend(fatTopPlot)
```

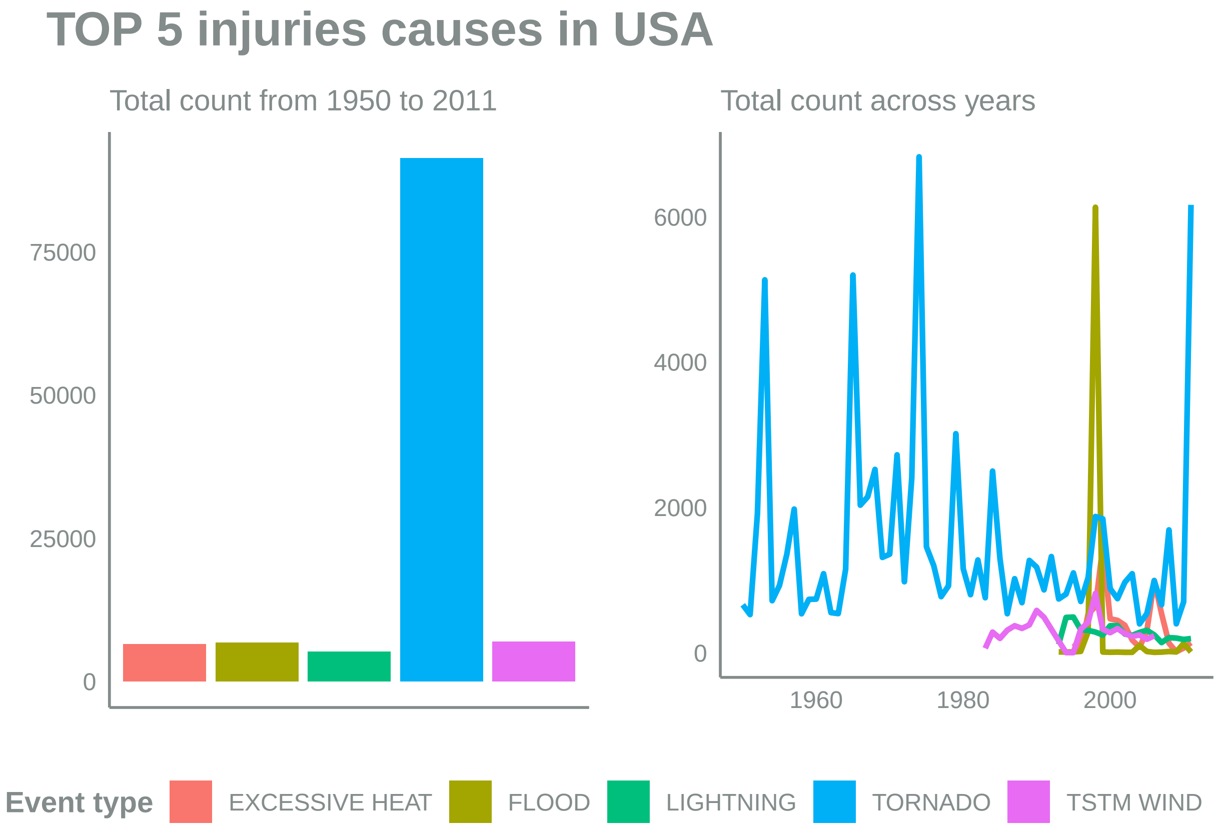
Then, the *grid* and *gridExtra* package were used to plot the graphics together.

```
grid.arrange(arrangeGrob(fatTopPlot + theme(legend.position = "none"),
                         fYearSeries + theme(legend.position = "none"),
                         nrow = 1),
             fatLegend,
             nrow = 2,
             heights = c(10, 1),
             top = textGrob(label = "TOP 5 fatalities causes in USA",
                            gp = gpar(col = "azure4", fontface = "bold", cex = 1.5),
                            hjust = -0.1,
                            vjust = 0.5,
                            x = 0,
                            y = 0))
```



4.2 Injuries

The same was done to **injuries**. For sake of simplification, the code will not be show.

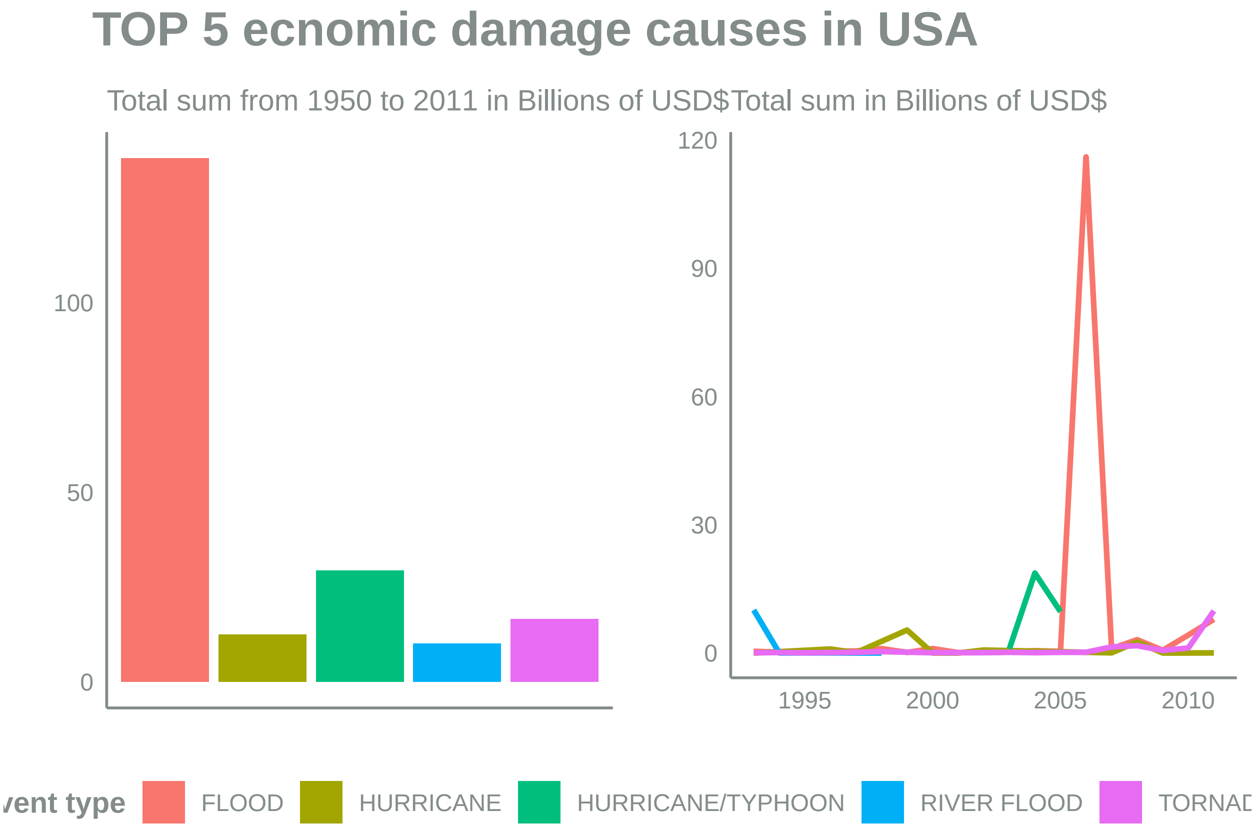


5 Analysing data from economic damage

For **economic** damage analysis, a sum of both property and crop damage was used.

```
# Top 5 economic damage causes totals

ecoTop <- economic %>%
        mutate(tDmg = (CROPDMG * CROPDMGEXP + PROPDMG * PROPDMGEXP)/1000000000) %>%
        group_by(EVTYPE) %>%
        summarise(dmg.total = sum(tDmg)) %>%
        top_n(n = 5, wt = dmg.total)

# Top 5 economic damage causes time series

ecoTime <- economic %>%
        mutate(tDmg = (CROPDMG * CROPDMGEXP + PROPDMG * PROPDMGEXP) / 1000000000) %>%
        group_by(EVTYPE, year(BGN_DATE)) %>%
        summarise(eco.total = sum(tDmg)) %>%
        filter(EVTYPE %in% ecoTop$EVTYPE) %>%
        rename(eco.year = `year(BGN_DATE)`)
```

The individual graphics was constructed the same way as **fatalities** and **injuries**. So, for the sake of simplicity, the code will not be show.



6 Conclusion

Related to question 01 (heath impact), it is difficult to say that tornados are the most frequent cause to fatalities, due to the fact that its data has been collected for a longer time than other causes. After 1990 other causes becomes important, like heat and excessive heat, while tornados fatalities have fallen down (maybe due to better predictions that allow people to get shelter).

The same behavior is observed to injuries, tornados occurrences fallen down, while flood increasing.

The only exception was 2011 year, when tornados overcame all other fatalities injuries. This happened because this year several tornadoes reach several states, mainly Alabama.

Related to second question, Food represents the most important cause of damage.