Between 1950 and November 2011 which weather events have had the largest impact on population health and property damage in the United States

Scott Robertson

02/10/2018

```
knitr::opts_chunk$set(echo = TRUE, warning = FALSE, message = FALSE, cache = TRUE, fig.path='figure/')
```

1. Synopsis

Storms and other severe weather events can cause both public health and economic problems for communities and municipalities. Many severe events can result in fatalities, injuries, and property damage, and preventing such outcomes to the extent possible is a key concern.

This project involves exploring the U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database. 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.

This data has been used to answer two questions in this paper:

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

---

2. Environment setup

During this analysis the following R packages will be used. The code provided below will check if they are already installed, install any missing ones, and load them into your working envrionment.

```
# Load the packages into the R environment
library(datasets)
library(dplyr)
library(ggplot2)
library(ggpubr)
```

The data for this project is avalible in the form of a comma-separated-value file compressed via the bzip2 algorithm to reduce its size. You can download the file from here:

- Storm Data [47Mb]

There is also some documentation of the database available. This provides details of how the raw data was captured and the structure of the data.

- National Weather Service Storm Data Documentation
- National Climatic Data Center Storm Events FAQ

The following code will create a data folder in your working directory, download the nessecary files and read the data into your environment.

```
# Create data folder in working directory
if (!file.exists("data/raw")) {
        dir.create("data/raw")
}

# Store url of data file as a variable
url1 <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
url2 <- "https://d396qusza40orc.cloudfront.net/repdata%2Fpeer2_doc%2Fpd01016005curr.pdf"
url3 <- "https://d396qusza40orc.cloudfront.net/repdata%2Fpeer2_doc%2FNCDC%20Storm%20Events-FAQ%20Page.p

# Download data file and supporting documents to data folder and store download time
download.file(url1, "./data/raw/storm_data.csv.bz2", method = "curl")
download.file(url2, "./data/raw/storm_data_documentation.pdf", method = "curl")
download.file(url3, "./data/raw/storm_data_faqs.pdf", method = "curl")
date_downloaded <- Sys.time()

# Output date data was downloaded
date_downloaded
```

```
## [1] "2018-10-08 16:57:30 BST"
```

---

3. Data processing

```
# Read full data set into R and show head of file
storm_data <- read.csv("./data/raw/storm_data.csv.bz2", header = TRUE, sep = ",")

# Look at the head of the data set to get idea of structure and identify fields needed for analysis
head(storm_data)
```

```
##   STATE__           BGN_DATE BGN_TIME TIME_ZONE COUNTY COUNTYNAME STATE
## 1       1  4/18/1950 0:00:00     0130       CST     97     MOBILE    AL
## 2       1  4/18/1950 0:00:00     0145       CST      3    BALDWIN    AL
## 3       1  2/20/1951 0:00:00     1600       CST     57    FAYETTE    AL
## 4       1   6/8/1951 0:00:00     0900       CST     89    MADISON    AL
## 5       1 11/15/1951 0:00:00     1500       CST     43    CULLMAN    AL
## 6       1 11/15/1951 0:00:00     2000       CST     77 LAUDERDALE    AL
##    EVTYPE BGN_RANGE BGN_AZI BGN_LOCATI END_DATE END_TIME COUNTY_END
## 1 TORNADO         0                                               0
## 2 TORNADO         0                                               0
## 3 TORNADO         0                                               0
## 4 TORNADO         0                                               0
## 5 TORNADO         0                                               0
## 6 TORNADO         0                                               0
##   COUNTYENDN END_RANGE END_AZI END_LOCATI LENGTH WIDTH F MAG FATALITIES
## 1         NA         0                      14.0   100 3   0          0
## 2         NA         0                       2.0   150 2   0          0
## 3         NA         0                       0.1   123 2   0          0
## 4         NA         0                       0.0   100 2   0          0
## 5         NA         0                       0.0   150 2   0          0
## 6         NA         0                       1.5   177 2   0          0
##   INJURIES PROPDMG PROPDMGEXP CROPDMG CROPDMGEXP WFO STATEOFFIC ZONENAMES
```

```
## 1       15    25.0          K       0
## 2        0     2.5          K       0
## 3        2    25.0          K       0
## 4        2     2.5          K       0
## 5        2     2.5          K       0
## 6        6     2.5          K       0
##   LATITUDE LONGITUDE LATITUDE_E LONGITUDE_ REMARKS REFNUM
## 1     3040      8812       3051       8806              1
## 2     3042      8755          0          0              2
## 3     3340      8742          0          0              3
## 4     3458      8626          0          0              4
## 5     3412      8642          0          0              5
## 6     3450      8748          0          0              6
```

```
# Check that the data set only contains information for the 50 USA states
str(storm_data$STATE)
```

```
##  Factor w/ 72 levels "AK","AL","AM",..: 2 2 2 2 2 2 2 2 2 2 ...
```

The current dataset contains a lot of fields which are not nessecary for our analysis. For our study data set we will create a subset with only the information related to event type, population health and property damage.

```
# Select only the columns of interest to our study
tidy_storm_data <- storm_data %>%
        select(STATE,
               EVTYPE,
               FATALITIES,
               INJURIES,
               PROPDMG,
               PROPDMGEXP,
               CROPDMG,
               CROPDMGEXP)
```

It also contains information for all United States territories, not just the core states. Before performing any analysis we will need to filter out any non-State entities.

```
# Filter the data using the state.abb field from the state file in datasets package
tidy_storm_data <- tidy_storm_data %>%
        filter(STATE %in% state.abb)

head(tidy_storm_data)
```

```
##   STATE  EVTYPE FATALITIES INJURIES PROPDMG PROPDMGEXP CROPDMG CROPDMGEXP
## 1    AL TORNADO          0       15    25.0          K       0
## 2    AL TORNADO          0        0     2.5          K       0
## 3    AL TORNADO          0        2    25.0          K       0
## 4    AL TORNADO          0        2     2.5          K       0
## 5    AL TORNADO          0        2     2.5          K       0
## 6    AL TORNADO          0        6     2.5          K       0
```

Now that we have only the fields and states that are interested in we need to turn change the property and crop damage totals into single unit representations.

In order to do this we use the following function to turn the PROPDMGEXP and CROPDMGEXP columns into numeric lists. We can then uses these lists to create cost columns by multiplying them by the PROPDMG and CROPDMG columns.

```
# Create a function that turns the EXP column into numeric values
exp_to_num <- function(i) {
    if (i %in% c("h", "H"))
        return(100)
    else if (i %in% c("k", "K"))
        return(1000)
    else if (i %in% c("m", "M"))
        return(1000000)
    else if (i %in% c("b", "B"))
        return(1000000000)
    else if (!is.na(as.numeric(i)))
        return(as.numeric(i))
    else if (i %in% c("", "-", "?", "+"))
        return(0)
}

# Convert property damage into cash value
prop_exp <- sapply(tidy_storm_data$PROPDMGEXP, FUN=exp_to_num)
tidy_storm_data$property_damage <- tidy_storm_data$PROPDMG * prop_exp

# Convert crop damage into cash value
crop_exp <- sapply(tidy_storm_data$CROPDMGEXP, FUN=exp_to_num)
tidy_storm_data$crop_damage <- tidy_storm_data$CROPDMG * crop_exp

# Drop unessecary property and crop columns
tidy_storm_data$PROPDMG <- NULL
tidy_storm_data$PROPDMGEXP <- NULL
tidy_storm_data$CROPDMG <- NULL
tidy_storm_data$CROPDMGEXP <- NULL
```

Finally we will rename the columns and save a copy of the tidy data set as an csv file in the data folder.

```
# Rename columns to make them easier to read
colnames(tidy_storm_data) <- c("state",
                               "event_type",
                               "fatalities",
                               "injuries",
                               "property_damage_dollars",
                               "crop_damage_dollars")

# Show head of tidy data
head(tidy_storm_data)
```

```
##   state event_type fatalities injuries property_damage_dollars
## 1    AL    TORNADO          0       15                   25000
## 2    AL    TORNADO          0        0                    2500
## 3    AL    TORNADO          0        2                   25000
## 4    AL    TORNADO          0        2                    2500
## 5    AL    TORNADO          0        2                    2500
## 6    AL    TORNADO          0        6                    2500
##   crop_damage_dollars
## 1                   0
## 2                   0
## 3                   0
## 4                   0
```

```
## 5                   0
## 6                   0
```

```
# Save a copy of the tidy data set to the data directory
if (!file.exists("data/tidy")) {
        dir.create("data/tidy")
}
write.csv(tidy_storm_data, file = "./data/tidy/tidy_storm_data.csv")
```

---

4. Population Health Impact Prepration

In order to assess which events have the highest impact on population health we will look at which events have caused the highest total fatalities and injuries.

The sum is being used as opposed to the average per event as this will be less likley to be negativley impacted by inconsistent recording of the event type over the length of the dataset, in particular in earlier periods when not all events were captured.

It also avoids focusing on extreme events which may have only occured a handful of times, with significant impact, and instead focuses on repeating events which can be targeted for future improvements in management.

```
# Summarise dataset by total fatalities and injuries for each event
health_impact <- tidy_storm_data %>%
        group_by(event_type) %>%
        summarise(fatalities = sum(fatalities), injuries = sum(injuries))
```

---

5. Economic Impact Preperation

For the same reason as mentioned in the Population Health Impact Prepration section we will be using the total cost in dollars for each event type.

As both property and crop damage are meassured in dollar ammounts we will also create a combined cost for each event type in case of any variation between damage types.

```
# Summarise dataset by total fatalities and injuries for each event
economic_impact <- tidy_storm_data %>%
        group_by(event_type) %>%
        summarise(property = sum(property_damage_dollars), crops = sum(crop_damage_dollars))

## Create a combined column in order to look at total damage
economic_impact$combined <- economic_impact$property + economic_impact$crops
```

---

6 Results

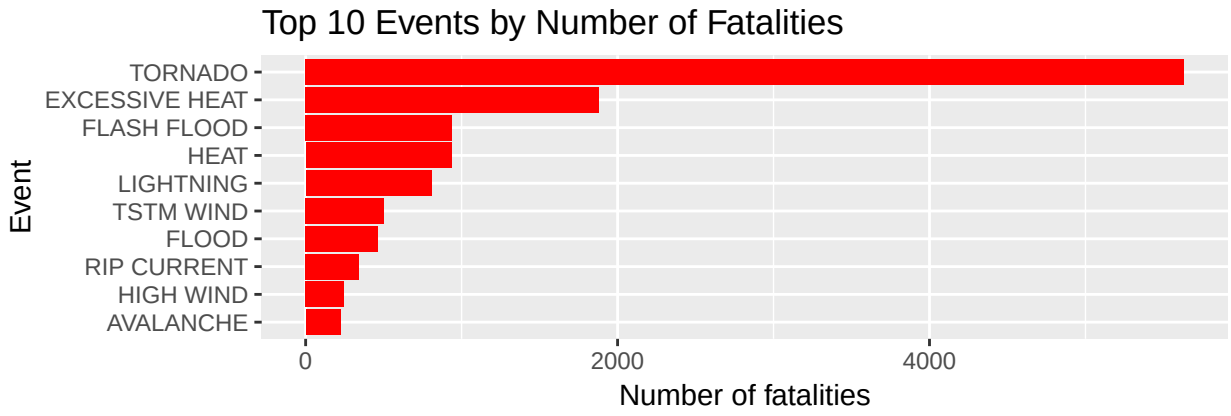
6.1 Across the United States, which types of events are most harmful with respect to population health?

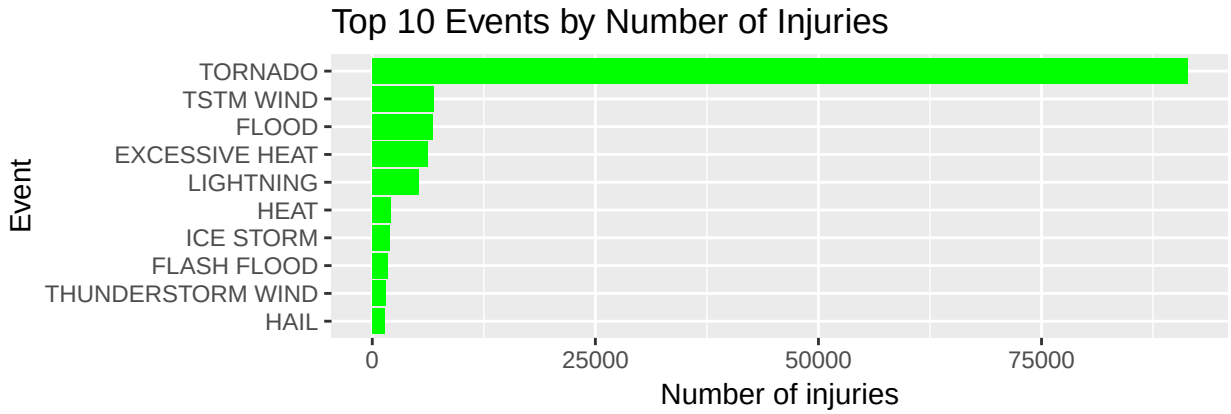
In order to look at the most harmful events the following figure has been produced showing the Top 10 events for number of Fatalities and Injuries.

```r
# Create two bar graphs to show the top 10 sources fatalities and injuries
fatalities <- ggplot(data=head(health_impact[order(health_impact$fatalities, decreasing = T),],10),
                     aes(x=reorder(event_type, fatalities), y=fatalities)) +
        geom_bar(fill="red",
                 stat="identity") +
        coord_flip() +
        labs (x = "Event",
              y = "Number of fatalities",
              title = "Top 10 Events by Number of Fatalities")

injury <- ggplot(data=head(health_impact[order(health_impact$injuries, decreasing = T),],10),
                 aes(x=reorder(event_type, injuries), y=injuries)) +
        geom_bar(fill="green",
                 stat="identity") +
        coord_flip() +
        labs (x = "Event",
              y = "Number of injuries",
              title = "Top 10 Events by Number of Injuries")

# Arrange the two graphs into a single figure
ggarrange(fatalities, injury, ncol = 1, nrow = 2)
```





Across both health metrics the events that have the greatest impact are Tornado's with **5633** fatalities and **91346** injuries between 1950 and November 2011.

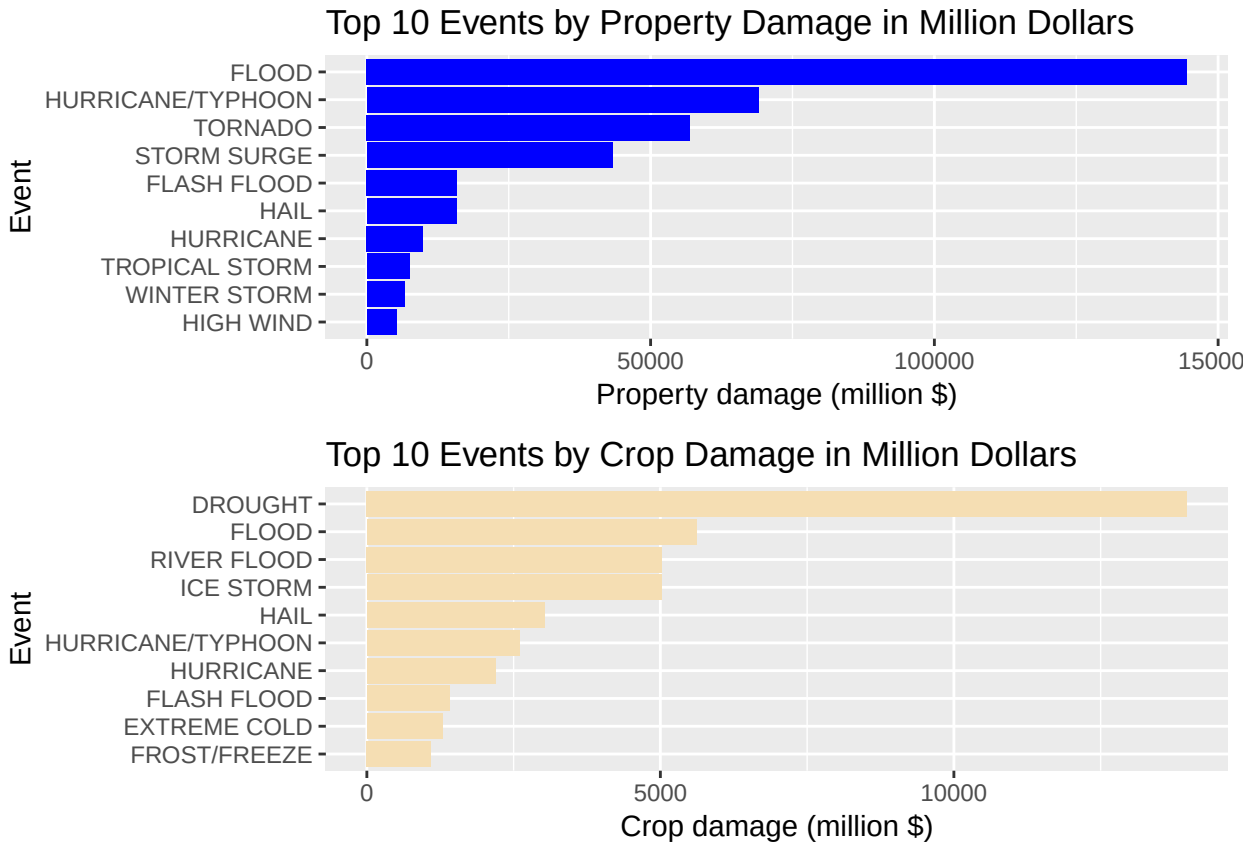
6.2 Across the United States, which types of events have the greatest economic consequences?

In order to answer this question we also look at the Top 10 events for Property Damage and Crop damage.

```
# Create two bar graphs to show the top 10 sources fatalities and injuries
property <- ggplot(data=head(economic_impact[order(economic_impact$property, decreasing = T),],10),
                   aes(x=reorder(event_type, property), y=property/1000000)) +
        geom_bar(fill="blue",
                 stat="identity") +
        coord_flip() +
        labs (x = "Event",
              y = "Property damage (million $)",
              title = "Top 10 Events by Property Damage in Million Dollars")

crops <- ggplot(data=head(economic_impact[order(economic_impact$crops, decreasing = T),],10),
                aes(x=reorder(event_type, crops), y=crops/1000000)) +
        geom_bar(fill="wheat",
                 stat="identity") +
        coord_flip() +
        labs (x = "Event",
              y = "Crop damage (million $)",
              title = "Top 10 Events by Crop Damage in Million Dollars")

# Arrange the two graphs into a single figure
ggarrange(property, crops, ncol = 1, nrow = 2)
```



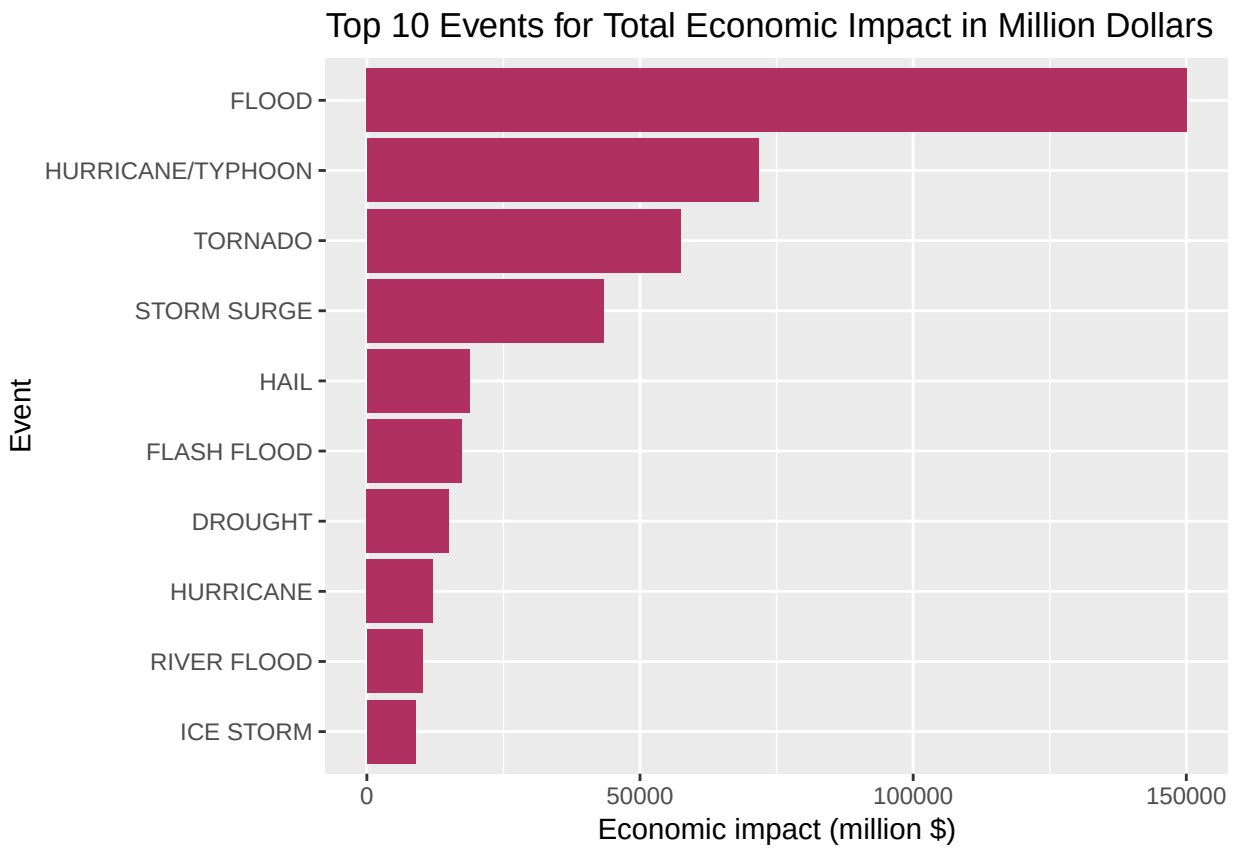
The image presented by these two graphs is less clear than with the health impact. The cost of the property damage is a lot higher than the crop damage, so you cannot easily tell from this figure which events have the

highest impact.

What we can draw from this is that Flood's have caused the most property damage at **$144531m** and Drought's have caused the most crop damage at **$13972m**.

We then combine the property and crop damage to create a total economic impact figure.

```
# Create two bar graphs to show the top 10 sources of combined damage cost
ggplot(data=head(economic_impact[order(economic_impact$combined, decreasing = T),],10),
       aes(x=reorder(event_type, combined), y=combined/1000000)) +
        geom_bar(fill="maroon",
                 stat="identity") +
        coord_flip() +
        labs(x = "Event",
             y = "Economic impact (million $)",
             title = "Top 10 Events for Total Economic Impact in Million Dollars")
```



From this it is possible to see that by a large margin the most economicly impactful weather event is flooding with a combined cost of **$150145m**.