Storm Data: Population Health and Economic Damages

Lindsay Spencer

April 10, 2016

Synopsis

Storms can be costly in a number of different ways. This research focuses on their cost to public health and the cost to the economy. The goal of this research is to find which storm events are the most costly in both areas. We find that tornados, excessive heat, and lightning have the highest impact to population health and floods and hurricanes have the highest economic costs.

Data Processing

The technical information

This research was performed on a CPU running Windows 10 Pro, with Intel(R) Core(TM) i7-4650U CPU @ 1.7GH and 2.3GHz.

Using R Studio Version 0.99.893 and R version 3.2.4 Revised.

The file was last downloaded on April 10, 2016 at 6:30 PM PST.

Load Data

First, we ensure that all the packages we will need are loaded and ready to process our data.

```
list.of.packages <- c("lubridate", "sqldf", "ggplot2", "gridExtra")
new.packages <-
    list.of.packages[!(list.of.packages %in% installed.packages()
                       [, "Package"])]

if (length(new.packages))
    install.packages(new.packages)

library(lubridate)
library(sqldf)
```

```
## Loading required package: gsubfn
```

```
## Loading required package: proto
## Loading required package: RSQLite
```

```
## Loading required package: DBI
```

```
library(ggplot2)
library(gridExtra)
```

Next, we download the storm data from the course website.

```
fileUrl <-
     "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
fileName = "stormData.csv"

if (!file.exists(fileName)) {
     download.file(fileUrl, fileName)
}
```

And we read in the file to the stormActivityData variable.

```
stormActivityData <- read.csv(fileName)
```

Data Cleaning

Subseting the Relevant Data

Let’s take a look at the raw data.

```
head(stormActivityData)
```

```
##   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
```

From this, we know we are only interested in a subset of the actual data. Specifically, we want the event type (EVTYPE), the number of fatalities (FATALITIES), the number of injuries (INJURIES), the property damage amount (PROPDMG and PROPDMGEXP), and the crop damage amount (CROPDMG and CROPDMGEXP). We will also use the date (BGN_DATE) to remove some of the data for dates that are probably incomplete.

```
relevantData <- c(
    "BGN_DATE",
    "EVTYPE",
    "FATALITIES",
    "INJURIES",
    "PROPDMG",
    "PROPDMGEXP",
    "CROPDMG",
    "CROPDMGEXP"
)

relevantDataSubset <- stormActivityData[relevantData]
```

Let’s take a look at this subsetted data.

```
head(relevantDataSubset)
```

```
##             BGN_DATE  EVTYPE FATALITIES INJURIES PROPDMG PROPDMGEXP
## 1  4/18/1950 0:00:00 TORNADO          0       15    25.0          K
## 2  4/18/1950 0:00:00 TORNADO          0        0     2.5          K
## 3  2/20/1951 0:00:00 TORNADO          0        2    25.0          K
## 4   6/8/1951 0:00:00 TORNADO          0        2     2.5          K
## 5 11/15/1951 0:00:00 TORNADO          0        2     2.5          K
## 6 11/15/1951 0:00:00 TORNADO          0        6     2.5          K
##   CROPDMG CROPDMGEXP
## 1       0
## 2       0
## 3       0
## 4       0
## 5       0
## 6       0
```

This is good, but we still need to do some transformations to be able to use this data for our analysis.

Caclulating Total Damages

The amount of damages to property and crops are presented as a number and an exponential factor. The documentation indicates that they use “K” for thousands, “M” for millions, and “B” for billions. The function, getMultiplier, was written to translate the EXP variable into a multiplier to calculate the correct amount of damages. It uses k and K, m and M, b and B, and ignores any other symbol that may be in the EXP column.

```
getMultiplier <- function(multiplier) {
     ifelse(
          multiplier == 'K' || multiplier == 'k',
          mult <- 1000,
          ifelse(
               multiplier == 'M' || multiplier == 'm',
               mult <- 1000000,
               ifelse(
                    multiplier == 'B' ||
                         multiplier == 'b',
                    mult <- 1000000000,
                    mult <- 1
               )
          )
     )
     mult
}
```

This function is now used to correctly record the amount of property damage and the amount of crop damage each event caused.

```
relevantDataSubset$propDamage <- relevantDataSubset$PROPDMG *
    as.integer(lapply(relevantDataSubset$PROPDMGEXP, getMultiplier))

relevantDataSubset$cropDamage <- relevantDataSubset$CROPDMG *
    as.integer(lapply(relevantDataSubset$CROPDMGEXP, getMultiplier))
```

Using these amounts, we calculate the total damage caused by each event.

```
relevantDataSubset$totalDamage <-
    relevantDataSubset$propDamage + relevantDataSubset$cropDamage
```

Here’s a look at what our dataset looks like after our calculations.

```
head(relevantDataSubset)
```

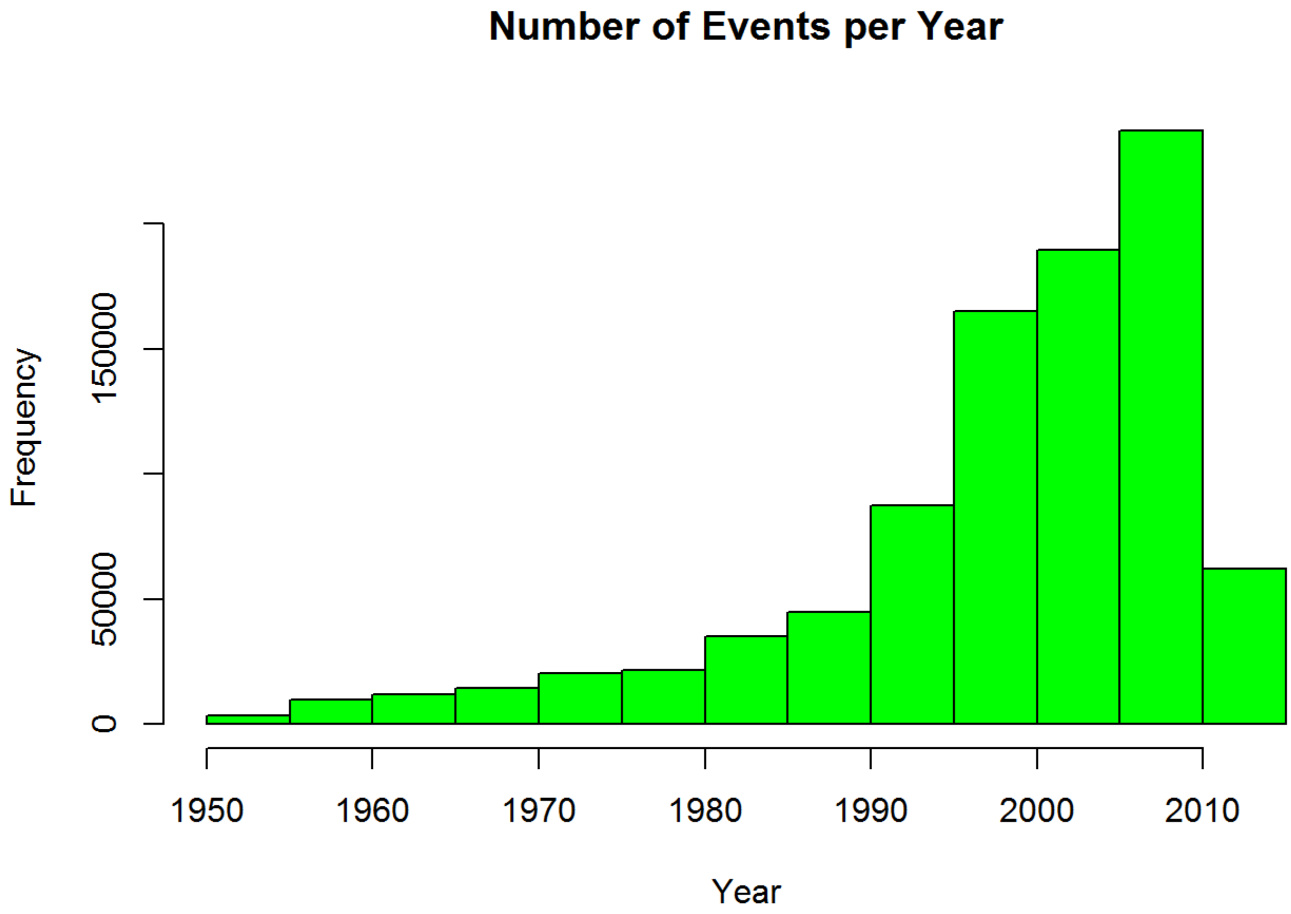
```
##             BGN_DATE  EVTYPE FATALITIES INJURIES PROPDMG PROPDMGEXP
## 1  4/18/1950 0:00:00 TORNADO          0       15    25.0          K
## 2  4/18/1950 0:00:00 TORNADO          0        0     2.5          K
## 3  2/20/1951 0:00:00 TORNADO          0        2    25.0          K
## 4   6/8/1951 0:00:00 TORNADO          0        2     2.5          K
## 5 11/15/1951 0:00:00 TORNADO          0        2     2.5          K
## 6 11/15/1951 0:00:00 TORNADO          0        6     2.5          K
##   CROPDMG CROPDMGEXP propDamage cropDamage totalDamage
## 1       0                 25000          0       25000
## 2       0                  2500          0        2500
## 3       0                 25000          0       25000
## 4       0                  2500          0        2500
## 5       0                  2500          0        2500
## 6       0                  2500          0        2500
```

Subseting Years that Have More Recorded Events

From the documentation, we know that the earlier years probably do not contain complete information. A look at the histogram of number of events by year shows us that this is probably true. To do this, we’ll have to extract the year from the date column.

```
relevantDataSubset$YEAR <-
    year(as.Date(relevantDataSubset$BGN_DATE, '%m/%d/%Y'))

hist(relevantDataSubset$YEAR, xlab = "Year",
     main = "Number of Events per Year", col = "green")
```



This shows us clearly that the earlier years do not have nearly the number of events recorded as the later years do. We make the assumption that the number of events recorded is at fault and not that the number of events has been increasing, so we will only look at the second half of our data. To do this, we will calculate the 50th percentile and subset the data to only look at events from the more recent years.

```
cutOffYear <- quantile(relevantDataSubset$YEAR, c(.5))
cutOffYear
```

```
##  50%
## 2002
```

We will use this value in the next section to exclude data from the first half of the given timeframe.

Aggregate Totals by Event

Now, we will total the amount of fatalities, injuries, and total damage done by each type of event. We will also use this opportunity to exclude the first 50% of the data and to rename the headers.

```
selectStatement <- paste(
    "SELECT ",
    "EVTYPE AS Event,",
    "SUM(FATALITIES) AS Fatalities,",
    "SUM(INJURIES) AS Injuries,",
    "SUM(TotalDamage) AS Damage",
    "FROM relevantDataSubset ",
    "WHERE YEAR >= '",
    cutOffYear,
    "'",
    "GROUP BY EVTYPE"
)

finalDataSet <- sqldf(selectStatement)
```

```
## Loading required package: tcltk
```

Here is a sample of the final dataset.

```
head(finalDataSet)
```

```
##                    Event Fatalities Injuries  Damage
## 1         ABNORMALLY DRY          0        0       0
## 2         ABNORMALLY WET          0        0       0
## 3 ASTRONOMICAL HIGH TIDE          0        0 9425000
## 4  ASTRONOMICAL LOW TIDE          0        0  320000
## 5              AVALANCHE        145      103 2722300
## 6              BLACK ICE          0        0       0
```

Results

Question 1

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

To answer this question, we will plot both the top 5 fatal events and top 5 injury-producing events separately.

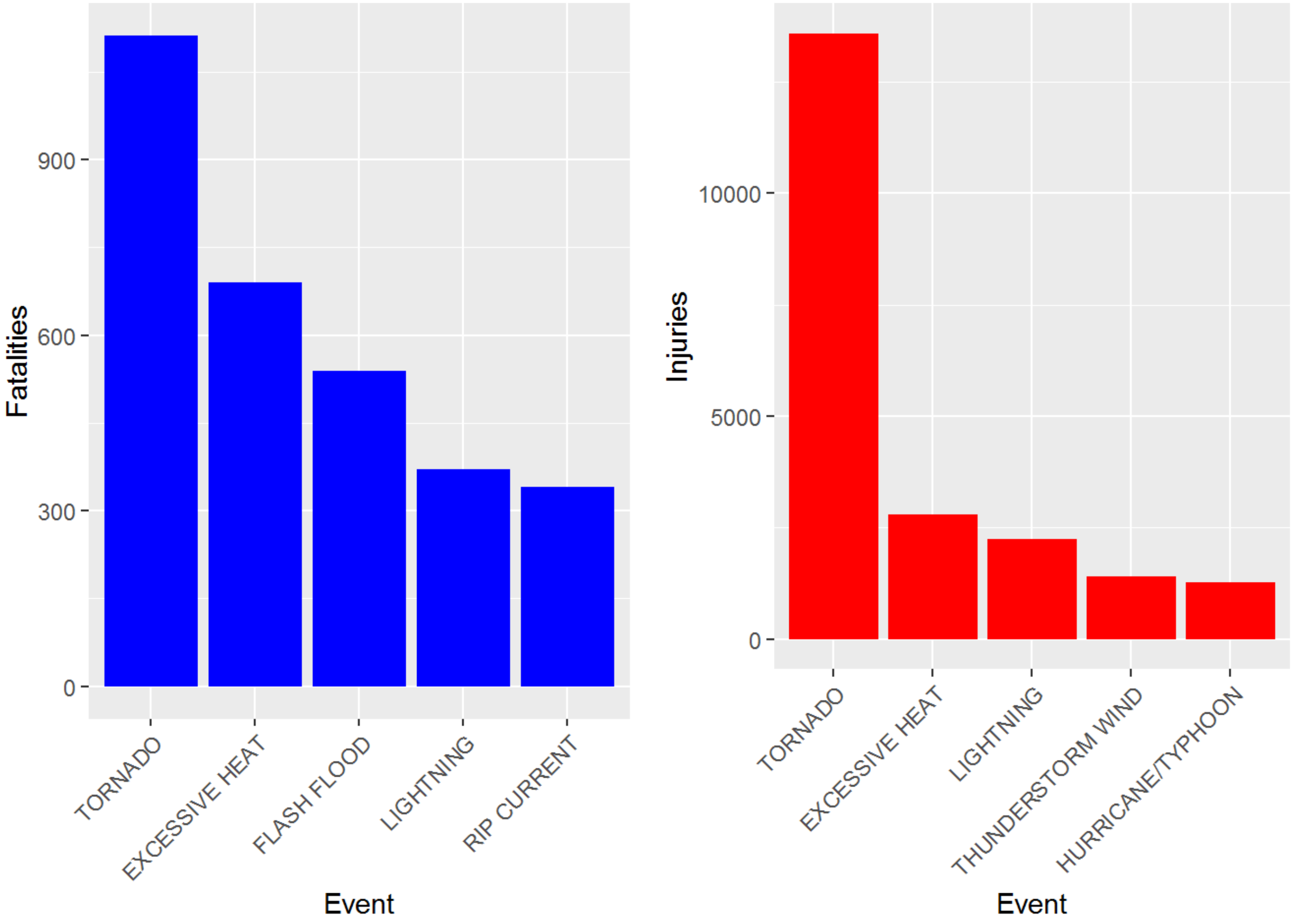
```
topFatalEvents <-
    head(finalDataSet[order(-finalDataSet$Fatalities), ], 5)

g1 <-
    ggplot(topFatalEvents,
           aes(x = factor(Event, levels = unique(Event)), y = Fatalities)) +
    geom_bar(stat = "identity", fill = "blue") +
    theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
    labs(x = "Event")

topInjuryEvents <-
    head(finalDataSet[order(-finalDataSet$Injuries), ], 5)

g2 <-
    ggplot(topInjuryEvents,
           aes(x = factor(Event, levels = unique(Event)), y = Injuries)) +
    geom_bar(stat = "identity", fill = "red") +
    theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
    labs(x = "Event")

grid.arrange(g1, g2, ncol = 2)
```



We can see that tornados cause both the greatest amount of fatalities and injuries, and that excessive heat causes the second most in both categories. Lightning is in the top 5 for both, as well. We can conclude that tornados, excessive heat, and lightning are the most harmful events with respect to public health.

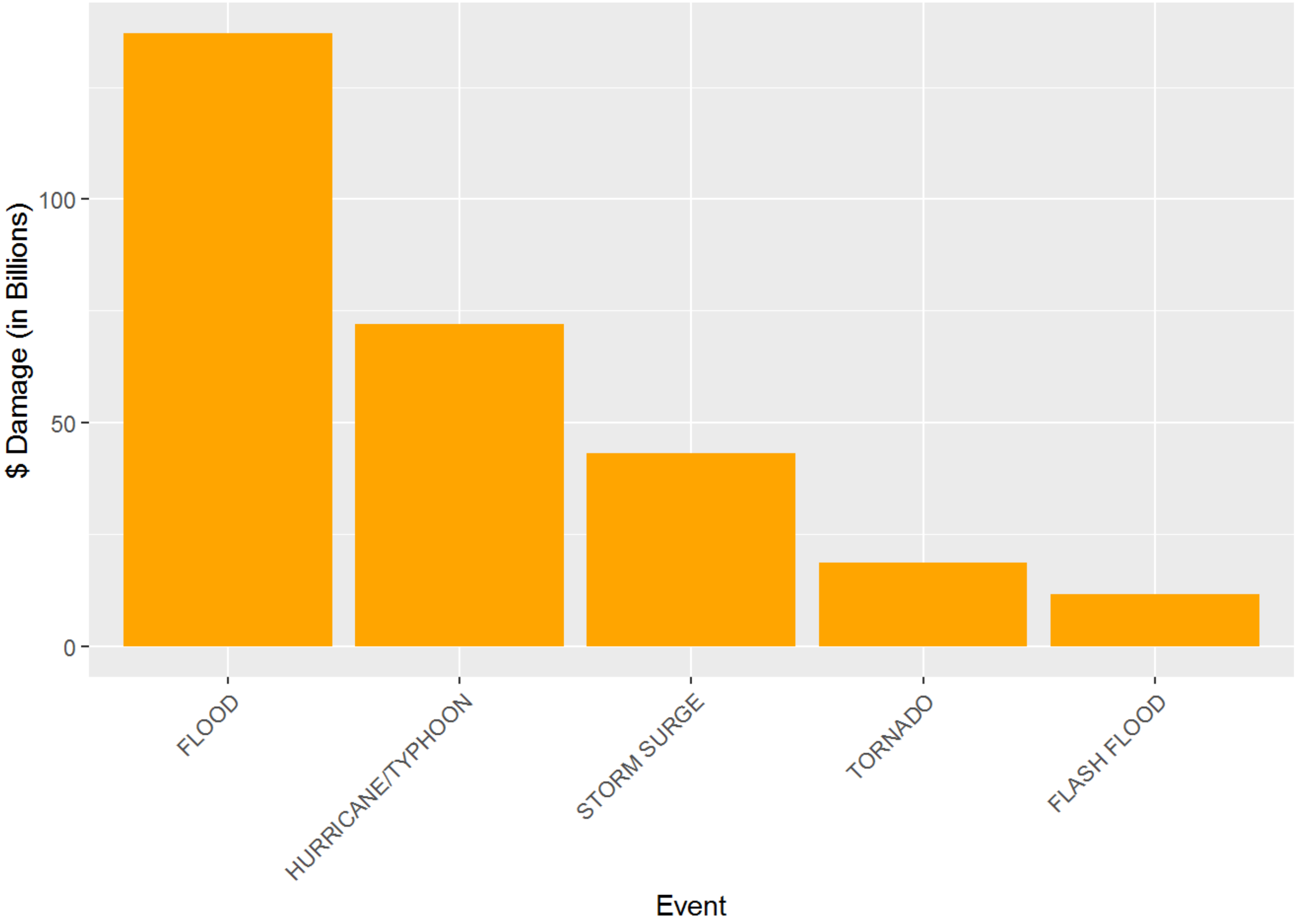
Question 2

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

For this question, we’ll plot the top 5 events with the greatest total economic damage.

```
economicDamageEvents <-
    head(finalDataSet[order(-finalDataSet$Damage),], 5)

ggplot(economicDamageEvents,
       aes(x = factor(Event, levels = unique(Event)), y = Damage)) +
    geom_bar(stat = "identity", fill = "orange") +
    theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
    labs(x = "Event", y = "$ Damage (in Billions)") +
    scale_y_continuous(
        labels = function(n) {
            format(n / 1000000000, scientific = FALSE)
        }
    )
```



From this graph, we see that floods cause the greatest economic damage, with hurricanes and typhoons coming in second.

Conclusion

Our analysis has shown us that tornados, excessive heat, and lightning are the most harmful events with respect to public health; while floods, hurricanes, and typhoons have the highest economic consequences.