Reproducible Research - Analyzing Storms

Nick Rumbaugh

March 20, 2018

Analysis of the Impact of Weather Events on Public Health and the Economy

Synopsis

We analyzed the impact of 985 different weather events on public health and the economy. We used two metrics each to measure impacts in these two areas: fatalities and injuries for public health, and property damage and crop damage for the economy. We found that tornados were the largest threat to both public health and the economy, having the highest total amoung of injuries, fatalities, and property damage. Several other events were dangerous for more specific reasons. Heat waves had the highest average number of injuries per event, hail was the largest damager of crops, and floods caused the highest average property damage per event.

Data Processing

First, we need to load in the data. The code below does that assuming the StormData.csv file is in the current directory. We use the data.table library for aggregate functions, so we load that and convert the data file to a data.table format.

```
df <- read.csv('./StormData.csv')
library(data.table)
dt <- data.table(df)
```

For this analysis, we are interested in five fields in the file (EVTYPE, FATALITIES, INJURIES, PROPDMG, CROPDMG). Before beginning the analysis, we need to check that these have no missing values.

```
for(name in c('EVTYPE','FATALITIES','INJURIES','PROPDMG','CROPDMG')) {
  print(sum(is.na(df[name])))
}
```

```
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
```

There are no missing values, so we are free to move on. Looking at the names of the weather events, there appear to be a large number of very similar events with slightly different names. For example, there are entries for excessive heat, extreme heat, heat wave, and heat wave drought. A more detailed analysis might try to combine some of these into categories, but that is beyond the scope of this relatively simple project, considering there are 985 unique values in EVTYPE. ### Public Health Analysis

First, we will address which weather events have the largest public health impact. There are a number of ways to approach this problem, but we will address just two metrics here: fatalities and injuries. Additionally, one could be interested in either which weather events have the largest total value of these metrics, or which ones have the largest average value.

A dangerous type of weather events may have high average and total injuries and fatalaties, but this may not always be the case. A very dangerous, but rare, weather event will have a high average value for the metrics, but a low total. Conversely, A very frequent, but less dangerous, event may have a high total value for the

metrics, but a low average. Now, let's calculate these metrics and look at the top 10 weather events for each one.

```
total_fatalities = dt[,list(total=sum(FATALITIES)),by=EVTYPE]
total_injuries = dt[,list(total=sum(INJURIES)),by=EVTYPE]
avg_fatalities = dt[,list(avg=mean(FATALITIES)),by=EVTYPE]
avg_injuries = dt[,list(avg=mean(INJURIES)),by=EVTYPE]
```

```
print(head(total_fatalities[order(total_fatalities$total, decreasing = TRUE),], 10 ))
```

```
##             EVTYPE total
##  1:        TORNADO  5633
##  2: EXCESSIVE HEAT  1903
##  3:    FLASH FLOOD   978
##  4:           HEAT   937
##  5:      LIGHTNING   816
##  6:      TSTM WIND   504
##  7:          FLOOD   470
##  8:    RIP CURRENT   368
##  9:      HIGH WIND   248
## 10:      AVALANCHE   224
```

```
print(head(avg_fatalities[order(avg_fatalities$avg, decreasing = TRUE),], 10))
```

```
##                         EVTYPE       avg
##  1: TORNADOES, TSTM WIND, HAIL 25.000000
##  2:              COLD AND SNOW 14.000000
##  3:      TROPICAL STORM GORDON  8.000000
##  4:      RECORD/EXCESSIVE HEAT  5.666667
##  5:               EXTREME HEAT  4.363636
##  6:             HIGH WIND/SEAS  4.000000
##  7:          HEAT WAVE DROUGHT  4.000000
##  8:              MARINE MISHAP  3.500000
##  9:              WINTER STORMS  3.333333
## 10:         HIGH WIND AND SEAS  3.000000
```

```
print(head(total_injuries[order(total_injuries$total, decreasing = TRUE),], 10))
```

```
##               EVTYPE total
##  1:           TORNADO 91346
##  2:         TSTM WIND  6957
##  3:             FLOOD  6789
##  4:    EXCESSIVE HEAT  6525
##  5:         LIGHTNING  5230
##  6:              HEAT  2100
##  7:         ICE STORM  1975
##  8:       FLASH FLOOD  1777
##  9: THUNDERSTORM WIND  1488
## 10:              HAIL  1361
```

```
print(head(avg_injuries[order(avg_injuries$avg, decreasing = TRUE),], 10))
```

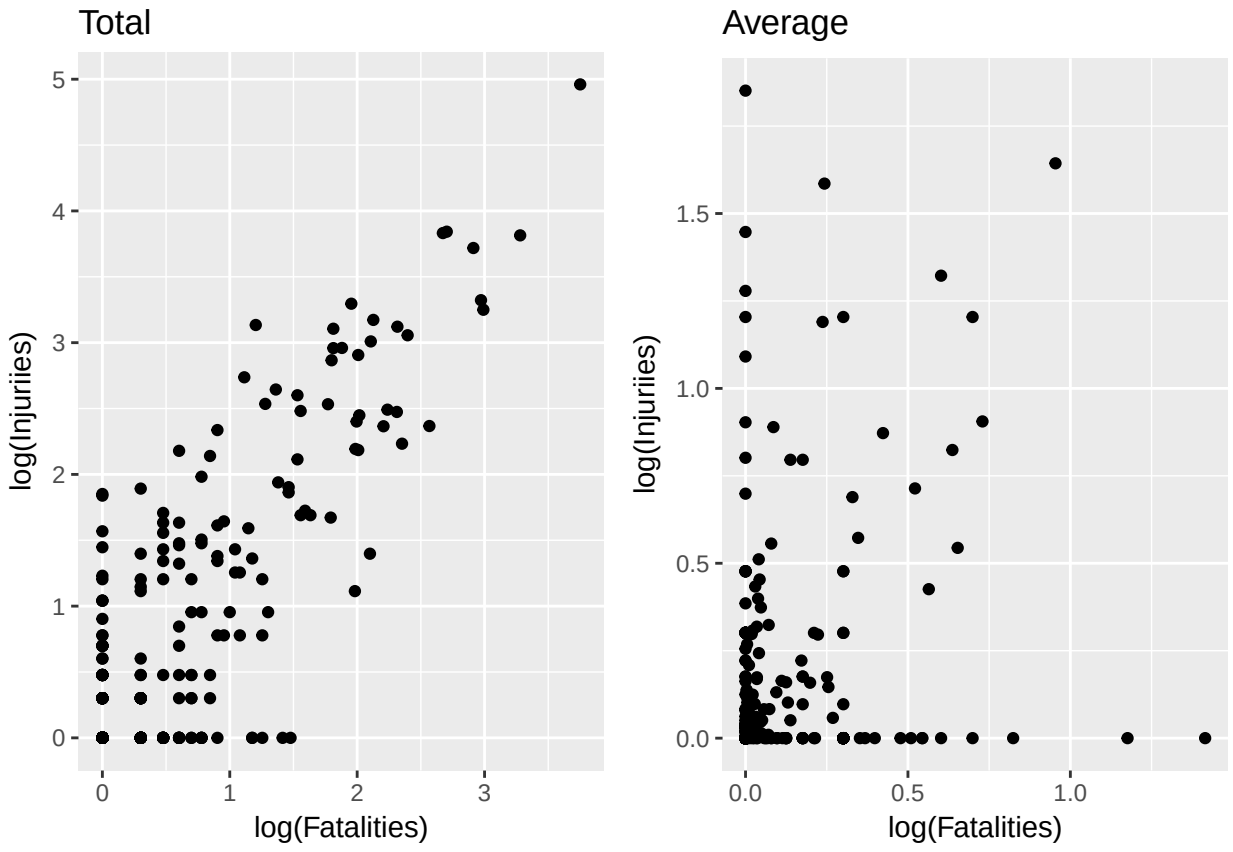
```
##                      EVTYPE      avg
##  1:              Heat Wave 70.00000
##  2:   TROPICAL STORM GORDON 43.00000
##  3:             WILD FIRES 37.50000
##  4:           THUNDERSTORMW 27.00000
```

```
##  5:      HIGH WIND AND SEAS 20.00000
##  6:         SNOW/HIGH WINDS 18.00000
##  7: WINTER STORM HIGH WINDS 15.00000
##  8:         GLAZE/ICE STORM 15.00000
##  9:       HEAT WAVE DROUGHT 15.00000
## 10:       HURRICANE/TYPHOON 14.48864
```

Tornadoes have, by far, the highest total injuries and fatalities. A tornado-related entry also has the highest average value for fatalaties. However, they are not in the top 10 for highest average injuries, where heat waves are the highest.

We can look at overall trends by plotting up the injury and fatality metrics we calculated for all types of weather events. To do this, we will use the ggplot2 and gridExtra libraries. We'll also plot in log space (after adding 1 to each value to avoid taking the log of 0) since there is a wide range of values.

```
library(ggplot2)
library(gridExtra)
total_df <- cbind(total_fatalities, total_injuries$total)
total_df$logfatalities <- log10(total_df$total + 1)
total_df$loginjuries <- log10(total_df$V2 + 1)
p1 <- qplot(logfatalities, loginjuries, data=total_df, main = 'Total',
            xlab = 'log(Fatalities)', ylab = 'log(Injuriies)')
avg_df <- cbind(avg_fatalities, avg_injuries$avg)
avg_df$logfatalities <- log10(avg_df$avg + 1)
avg_df$loginjuries <- log10(avg_df$V2 + 1)
p2 <- qplot(logfatalities, loginjuries, data=avg_df, main = 'Average',
            xlab = 'log(Fatalities)', ylab = 'log(Injuriies)')
grid.arrange(p1, p2, nrow=1)
```



The left plot shows that total fatalities and total injuries are highly correlated. Weather events that have high total fatalities also have high total injuries. Looking at averages in the right plot, this isn't necessarily the case. There are a number of events with high average injuries, without having high average fatalities. We'll discuss the tradeoffs of these different metrics in the Results section. ### Economic Analysis

In examining the economic impact of weather events, we will focus on two metrics: property damage (PROPDMG) and crop damage (CROPDMG), both of which are given in dollars in our data. As with the public health analysis, we will look at both total and average values.

```
total_propdmg = dt[,list(total=sum(PROPDMG)),by=EVTYPE]
total_cropdmg = dt[,list(total=sum(CROPDMG)),by=EVTYPE]
avg_propdmg = dt[,list(avg=mean(PROPDMG)),by=EVTYPE]
avg_cropdmg = dt[,list(avg=mean(CROPDMG)),by=EVTYPE]
```

```
print(head(total_propdmg[order(total_propdmg$total, decreasing = TRUE),], 10 ))
```

```
##                 EVTYPE     total
##  1:            TORNADO 3212258.2
##  2:        FLASH FLOOD 1420124.6
##  3:          TSTM WIND 1335965.6
##  4:              FLOOD  899938.5
##  5:  THUNDERSTORM WIND  876844.2
##  6:               HAIL  688693.4
##  7:          LIGHTNING  603351.8
##  8: THUNDERSTORM WINDS  446293.2
##  9:          HIGH WIND  324731.6
## 10:       WINTER STORM  132720.6
```

```
print(head(total_cropdmg[order(total_cropdmg$total, decreasing = TRUE),], 10 ))
```

```
##                 EVTYPE     total
##  1:               HAIL 579596.28
##  2:        FLASH FLOOD 179200.46
##  3:              FLOOD 168037.88
##  4:          TSTM WIND 109202.60
##  5:            TORNADO 100018.52
##  6:  THUNDERSTORM WIND  66791.45
##  7:            DROUGHT  33898.62
##  8: THUNDERSTORM WINDS  18684.93
##  9:          HIGH WIND  17283.21
## 10:         HEAVY RAIN  11122.80
```

```
print(head(avg_propdmg[order(avg_propdmg$avg, decreasing = TRUE),], 10 ))
```

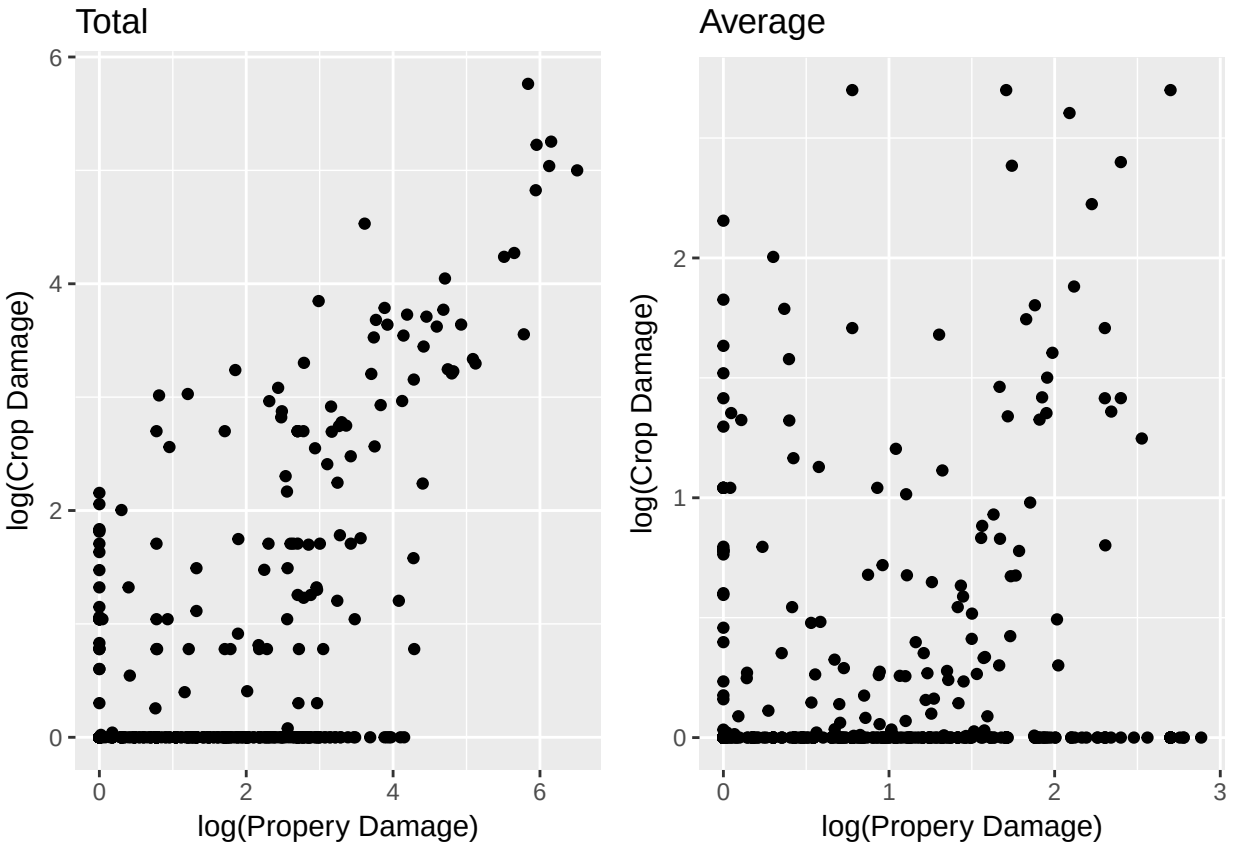
```
##                           EVTYPE avg
##  1:               COASTAL EROSION 766
##  2:         RIVER AND STREAM FLOOD 600
##  3:           HEAVY RAIN AND FLOOD 600
##  4:                      Landslump 570
##  5: FLASH FLOODING/THUNDERSTORM WI 500
##  6:                HIGH WIND/SEAS 500
##  7:                  FLASH FLOOD/ 500
##  8:          TROPICAL STORM GORDON 500
##  9:               HEAVY RAIN/SNOW 500
## 10:               SLEET/ICE STORM 500
```

```
print(head(avg_cropdmg[order(avg_cropdmg$avg, decreasing = TRUE),], 10 ))
```

```
##                    EVTYPE      avg
##  1: TROPICAL STORM GORDON 500.0000
##  2: DUST STORM/HIGH WINDS 500.0000
##  3:          FOREST FIRES 500.0000
##  4:       HIGH WINDS/COLD 401.0000
##  5:       HURRICANE FELIX 250.0000
##  6:        River Flooding 241.3680
##  7:         WINTER STORMS 166.6667
##  8:     EXCESSIVE WETNESS 142.0000
##  9:          Frost/Freeze 100.0000
## 10:               TYPHOON  75.0000
```

Looking at the total property and crop damage, we can see that tornados are also one of the top damages. They cause the most property damage and are in the top 10 for crop damage. Hail wasn't a large public health threat, but it is the top crop damager and is in the top 10 for property damage. Looking at average damage values doesn't appear to be as useful here. Flooding-type events, which were also in the top 10 for total property damage, make up most of the top 10 average property damagers. The top 10 average crop damagers have some specific tropical storms and hurricanes. In a more detailed analysis, these could be grouped into one category. Again, we also want to look at plots to see the overall trends.

```
total_df <- cbind(total_propdmg, total_cropdmg$total)
total_df$logpropdmg <- log10(total_df$total + 1)
total_df$logcropdmg <- log10(total_df$V2 + 1)
p1 <- qplot(logpropdmg, logcropdmg, data=total_df, main = 'Total',
            xlab = 'log(Propery Damage)', ylab = 'log(Crop Damage)')
avg_df <- cbind(avg_propdmg, avg_cropdmg$avg)
avg_df$logpropdmg <- log10(avg_df$avg + 1)
avg_df$logcropdmg <- log10(avg_df$V2 + 1)
p2 <- qplot(logpropdmg, logcropdmg, data=avg_df, main = 'Average',
            xlab = 'log(Propery Damage)', ylab = 'log(Crop Damage)')
grid.arrange(p1, p2, nrow=1)
```



For both total and average values, property and crop damage appear to be highly correlated. However, there do appear to be a lot of weather events that cause lots of property damage, but have no recorded crop damage.

Results

The goal of this analysis was to find the weather events that have the highest impact on public health and the economy. The overall highest threat appears to be tornados. Tornados cause the most total injuries, deaths, and property damage of any weather event, and cause the fifth most crop damage. Looking at average values, tornados are not as damaging. This is likely because they cause damage in a very narrow path, which can easily miss people or property and there are a relatively large number of them. However, when they do hit, they do a lot of damage. The relatively low average value of injuries from tornados could also indicate that they tend to cause deaths compared to injuries at a higher rate compared to other types of events.

There are other events that are dangerous to public health in different ways. Heat waves cause the most number of injuries, on average, of any event. This is likely because these have a large area of effect and it is very difficult to escape their impact. They may not cause as many injuries or deaths as tornados, but they are reliably dangerous when they occur.

In terms of economic impact, while tornados are still one of the predominant threats, there are others that are also dangerous in different ways. Hail doesn't tend to cause injuries or deaths in the way tornados does, and does not cause as much property damage, but it causes the most crop damage by a factor of 3. If one is concerned with impact on the agricultural industry, hail should be the top priority.

Floods are also one of the top property and crop damagers, and have the highest average property damage per event. Lik heat waves, this is likely because these events have large area of effect and it is difficult to escape their impact.

There were several shortcoming of this analysis. Primarily, the weather events tended to be categorized into

similar, but differently named, categories. For example, there were separate categories for excessive heat, extreme heat, and heat waves. In addition, there were categories for individual tropical storms and hurricanes. In a more detailed analysis, these events would be grouped together, but the size of the dataset (there are 985 unique entries for the EVTYPE variable) precludes that for this simple analysis. This shortcoming is particularly relevant for hurricanes, which likely have a much higher impact than suggested by this analysis.