Title: Reproducible Research: Peer Assessment Two

Dr Paul Fergus

24/10/2014

Summary

It is common knowledge that stroms and severe weather conditions have a serious impact on public health and the economy. There is a general consensus that the utilisation of data will help to inform and prioritise decisions about if and when they could occur, including estimates of potential fatalities, injury and damage to proporty. This report draws the readers attention to this fact and presents quantitative evidence based on a storm database collected from the U.S. Oceanic and Atmospheric Administration (NOAA) from 1950-2011.

Synopsis

The results in this report present us with some interesting results. From the first set of results we find that tornados have accounted for the number of deaths since recordings began in 1950, with 5633 recorded fatalities. This was followed by excessive heat wih 1903 recordings and the lowest end of the scale 224 of deaths caused by Avalanche. In terms of injury again we find that tornados account for most recorded insidences with 91346 recordings since 1950. At the lower end of the top ten events we are interesed in in this report we find that 1361 injuries were recorded. Looking at the economic burden severe weather conditions has on the economy we find that floods account for the main cause of damage to property at a total cost of 115 billion since 1950 with hurricane/typhoon, and storm surge accounding for 58 billion and 31 billion respectively. (the results here are problematic because of duplication in the records). In terms of the costs cuased by damage to crops we find that river floods and ice storms acount for the most economic costs, costing 5, 5, and 1.5 billion respectively.

Load Libraries and Perform Initial Setup

```
#Load Libraries
library(ggplot2)
library(plyr)
#This allows the reader to examine the software environment
sessionInfo()
```

```
## R version 3.0.2 (2013-09-25)
## Platform: x86_64-pc-linux-gnu (64-bit)
##
## locale:
##  [1] LC_CTYPE=en_GB.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_GB.UTF-8        LC_COLLATE=en_GB.UTF-8
##  [5] LC_MONETARY=en_GB.UTF-8    LC_MESSAGES=en_GB.UTF-8
##  [7] LC_PAPER=en_GB.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_GB.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] plyr_1.8.1    ggplot2_1.0.0
##
## loaded via a namespace (and not attached):
##  [1] colorspace_1.2-4 digest_0.6.4     evaluate_0.5.5   formatR_1.0
##  [5] grid_3.0.2       gtable_0.1.2     htmltools_0.2.6  knitr_1.7
##  [9] MASS_7.3-29      munsell_0.4.2    proto_0.3-10     Rcpp_0.11.2
## [13] reshape2_1.4     rmarkdown_0.2.68 scales_0.2.4     stringr_0.6.2
## [17] tools_3.0.2      yaml_2.1.13
```

Load Data

```
if(!file.exists("./data/repdata-data-StormData.csv")){
  dataset <- read.table(
    bzfile("./data/repdata-data-StormData.csv.bz2",
      "repdata-data-StormData.csv"), sep=",", header=T, na.string="NA")
}
```

Check that we have data

At this stage we just want to check that we have loaded some data.

```
head(dataset, n=1)
```

```
##   STATE__           BGN_DATE BGN_TIME TIME_ZONE COUNTY COUNTYNAME
STATE
## 1       1 4/18/1950 0:00:00     0130       CST     97     MOBILE    AL
##    EVTYPE BGN_RANGE BGN_AZI BGN_LOCATI END_DATE END_TIME
COUNTY_END
## 1 TORNADO         0                                               0
##   COUNTYENDN END_RANGE END_AZI END_LOCATI LENGTH WIDTH F MAG
FATALITIES
## 1         NA         0                         14   100 3   0          0
##   INJURIES PROPDMG PROPDMGEXP CROPDMG CROPDMGEXP WFO
STATEOFFIC ZONENAMES
## 1       15      25          K       0
```

```
##   LATITUDE LONGITUDE LATITUDE_E LONGITUDE_ REMARKS REFNUM
## 1     3040      8812       3051       8806              1
```

Data Preprocessing

I left completing this assignment far too late. I would have liked to have cleaned the data more than I have. In particular, the final set of results caused problems becuase of dupiliate EVTYPE data values- please see the numeric values - I should have dealt with these values in the preprocessing stage, but have not due to time constraints.

What has been done is to convert the alpha values that represent thousands, millions and billions,into numerical equivalents so that we can calculate the property and crop damage. For this project I have capped the number of events that we are interested in to the top ten as these account for most of the damage and costs.

During this stage we also work out which weather event causes the most loss of life and injury and which of the weather events recorded has the most significant ecomonic impact on property and crops

```
#PROPDMGEXP
#Convert the exponential to a numeric value that we can work with.
dataset$PROPDMGEXP <- as.character(dataset$PROPDMGEXP)

## Warning: closing unused connection 5
## (./data/repdata-data-StormData.csv.bz2)

dataset$PROPDMGEXP[grep("K", dataset$PROPDMGEXP)] <- "1000"
dataset$PROPDMGEXP[grep("k", dataset$PROPDMGEXP)] <- "1000"
dataset$PROPDMGEXP[grep("M", dataset$PROPDMGEXP)] <- "1000000"
dataset$PROPDMGEXP[grep("m", dataset$PROPDMGEXP)] <- "1000000"
dataset$PROPDMGEXP[grep("B", dataset$PROPDMGEXP)] <- "1000000000"
dataset$PROPDMGEXP[grep("b", dataset$PROPDMGEXP)] <- "1000000000"
#Set all other characters to 1 - we consider this to be noise in the data.
to.be.one <- dataset$PROPDMGEXP %in% c("1000", "1000000",
"1000000000") == F
dataset$PROPDMGEXP[to.be.one == TRUE] <- "1"
#Change the variable to numeric so that we can perform the calculations
dataset$PROPDMGEXP <- as.numeric(dataset$PROPDMGEXP)

#CROPDMGEXP
#Convert the exponential to a numeric value that we can work with.
dataset$CROPDMGEXP <- as.character(dataset$CROPDMGEXP)
dataset$CROPDMGEXP[grep("K", dataset$CROPDMGEXP)] <- "1000"
dataset$CROPDMGEXP[grep("k", dataset$CROPDMGEXP)] <- "1000"
dataset$CROPDMGEXP[grep("M", dataset$CROPDMGEXP)] <- "1000000"
dataset$CROPDMGEXP[grep("m", dataset$CROPDMGEXP)] <- "1000000"
dataset$CROPDMGEXP[grep("B", dataset$CROPDMGEXP)] <- "1000000000"
dataset$CROPDMGEXP[grep("b", dataset$CROPDMGEXP)] <- "1000000000"
```

```r
#Set all other characters to 1 - we consider this to be noise in the data.
to.be.one <- dataset$CROPDMGEXP %in% c("1000", "1000000",
"1000000000") == F
dataset$CROPDMGEXP[to.be.one == TRUE] <- "1"
#Change the variable to numeric so that we can perform the calculations
dataset$CROPDMGEXP <- as.numeric(dataset$CROPDMGEXP)

#Calculate the costs
dataset$prop.damage <- dataset$PROPDMG * dataset$PROPDMGEXP
dataset$crop.damage <- dataset$CROPDMG * dataset$CROPDMGEXP

#Extract the top 10 most
top.ten.prop <- head(dataset[order(dataset$prop.damage,
decreasing=TRUE),], 10)
top.ten.crop <- head(dataset[order(dataset$crop.damage,
decreasing=TRUE),], 10)

#First of all summerise all the fatalities and injuries for all of the event types
in the dataset.
casulties <- ddply(dataset, .(EVTYPE), summarize,
                  fatalities = sum(FATALITIES),
                  injuries = sum(INJURIES))

#Get the top 10  event types that cause fatality.
top.ten.fatalities <- head(casulties[order(casulties$fatalities,
decreasing=TRUE), ], 10)

#Get the top 10  event types that cause fatality.
top.ten.injuries <- head(casulties[order(casulties$injuries, decreasing=TRUE),
], 10)
```

Results

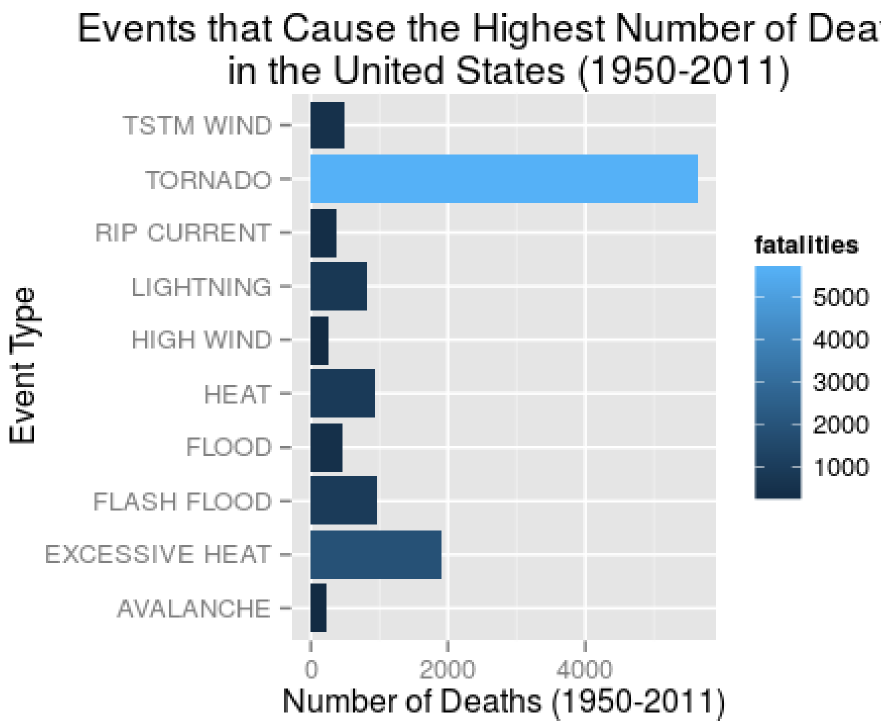
Question One: Across the United States, which types of events (as indicated in the EVTYPE variable) are most harmful with respect to population health

We are not interested in every event type, only the ones that cause the most harm. For the purposes of this study we are only interested in the top 10 Event Types that cause the most harm to the poluation (although this could be changed to suite specific needs).

The results below show the top ten number of fatalities and injuries categorised by event type

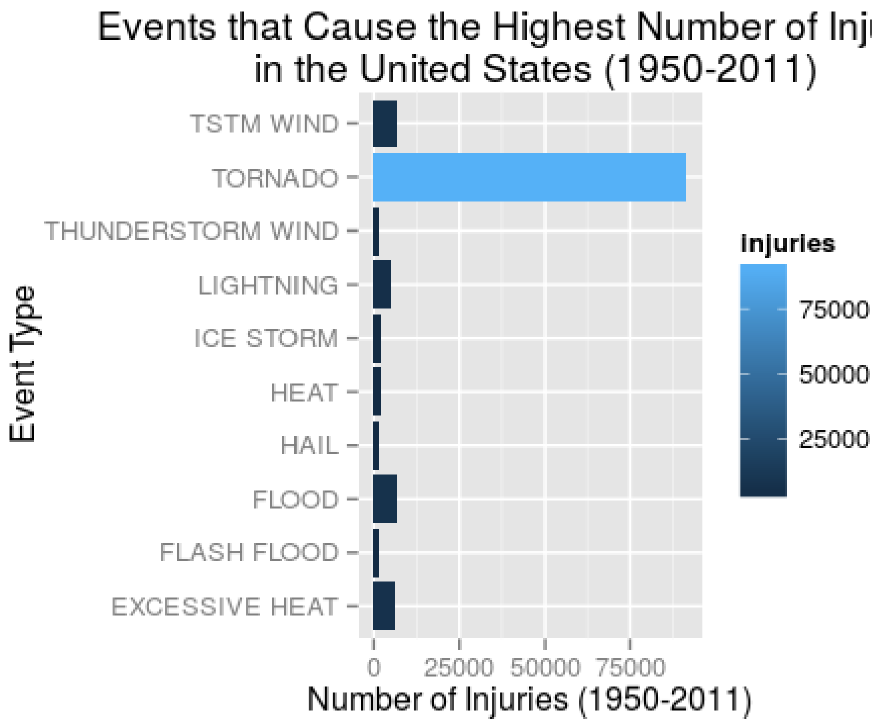
```r
##Top 10 Events that Caused the Highest Number of Deaths
#Plot top ten fatalities by event type
library(ggplot2)
ggplot(top.ten.fatalities, aes(EVTYPE, fatalities, fill=fatalities)) +
```

```
  geom_bar(stat="identity") + coord_flip() +
  stat_summary(fun.y = median, geom="bar") +
  labs(x="Event Type", y="Number of Deaths (1950-2011)",
       title="Events that Cause the Highest Number of Deaths\n in the United
States (1950-2011)")
```




```
#Display the top 10 fatalities
top.ten.fatalities[,c("EVTYPE", "fatalities")]

##              EVTYPE fatalities
## 830        TORNADO       5633
## 123 EXCESSIVE HEAT       1903
## 147    FLASH FLOOD        978
## 269           HEAT        937
## 452      LIGHTNING        816
## 854      TSTM WIND        504
## 164          FLOOD        470
## 581    RIP CURRENT        368
## 354      HIGH WIND        248
## 11       AVALANCHE        224

##Top 10 Events that Caused the Highest Number of Injuries
#Plot top ten injuries by event type
ggplot(top.ten.injuries, aes(EVTYPE, injuries, fill=injuries)) +
  geom_bar(stat="identity") + coord_flip() +
  stat_summary(fun.y = median, geom="bar") +
  labs(x="Event Type", y="Number of Injuries (1950-2011)",
```

```
    title="Events that Cause the Highest Number of Injuries \n in the United States (1950-2011)")
```



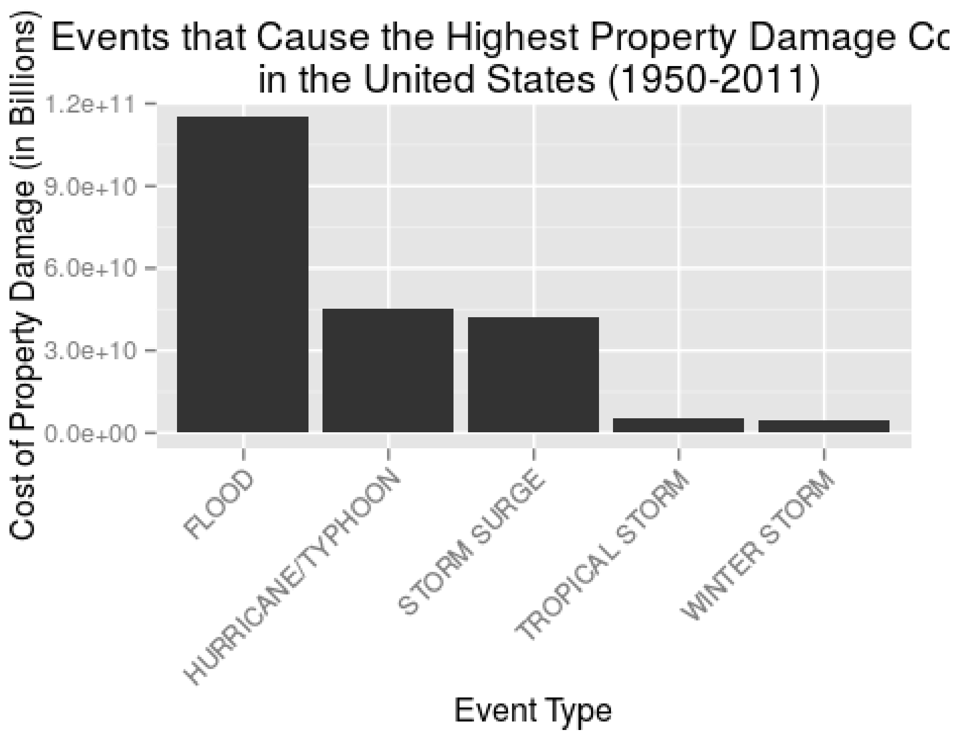

```
#Display the top 10 fatalities
top.ten.injuries[,c("EVTYPE", "injuries")]

##              EVTYPE injuries
## 830         TORNADO    91346
## 854       TSTM WIND     6957
## 164           FLOOD     6789
## 123  EXCESSIVE HEAT     6525
## 452       LIGHTNING     5230
## 269            HEAT     2100
## 424       ICE STORM     1975
## 147     FLASH FLOOD     1777
## 759 THUNDERSTORM WIND     1488
## 238            HAIL     1361
```

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

As with question one we are not interested in every event type, only the ones that have the greatest economical costs. For the purposes of this study we are only interested in the top 10 Event Types that cost the economoy the most (although this could be changed to suite specific needs).

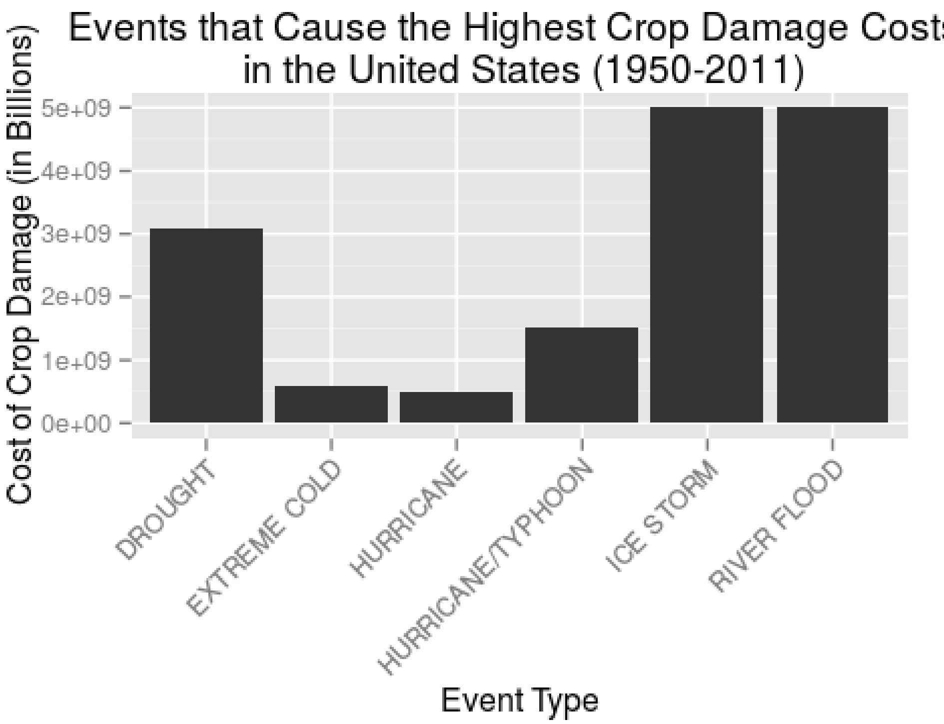
The results below show the top ten highest economical costs for prop and crop for the severest weather categorised by event type.

```r
#Plot the top 10 economoic prop damages
ggplot(top.ten.prop, aes(EVTYPE, prop.damage)) +
  geom_bar(stat="identity") +
  stat_summary(fun.y = median, geom="bar") +
  labs(x="Event Type", y="Cost of Property Damage (in Billions)",
       title="Events that Cause the Highest Property Damage Costs \n in the United States (1950-2011)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```r
#Display the top 10 prop damage
top.ten.prop[,c("EVTYPE", "prop.damage")]

##                  EVTYPE prop.damage
## 605953            FLOOD   1.150e+11
## 577676      STORM SURGE   3.130e+10
## 577675 HURRICANE/TYPHOON   1.693e+10
## 581535      STORM SURGE   1.126e+10
## 569308 HURRICANE/TYPHOON   1.000e+10
## 581533 HURRICANE/TYPHOON   7.350e+09
## 581537 HURRICANE/TYPHOON   5.880e+09
## 529351 HURRICANE/TYPHOON   5.420e+09
## 443782    TROPICAL STORM   5.150e+09
## 187564      WINTER STORM   5.000e+09
```

```r
#Plot the top 10 economoic crop damages
ggplot(top.ten.crop, aes(EVTYPE, crop.damage)) +
  geom_bar(stat="identity") +
  stat_summary(fun.y = median, geom="bar") +
  labs(x="Event Type", y="Cost of Crop Damage (in Billions)",
       title="Events that Cause the Highest Crop Damage Costs \n in the United States (1950-2011)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```r
#Display the top 10 crop damage
top.ten.crop[,c("EVTYPE", "crop.damage")]
```

```
##                   EVTYPE crop.damage
## 198389       RIVER FLOOD  5000000000
## 211900         ICE STORM  5000000000
## 581537 HURRICANE/TYPHOON  1510000000
## 639347           DROUGHT  1000000000
## 312986      EXTREME COLD   596000000
## 422676           DROUGHT   578850000
## 410175           DROUGHT   515000000
## 199733           DROUGHT   500000000
## 337008           DROUGHT   500000000
## 366694         HURRICANE   500000000
```