Storms-Ecomomic And Public Health Impact In The USA.

An exploration of NOAA storm data for the cost analysis of the differents types of storms.

Patrick Leugue

iNTELLIGENTIANALYTICs
(mailto:#)patrickleugue@gmail.com (mailto:patrickleugue@gmail.com)

Monday, August 19, 2014

Introduction

The United States are affected every year by natural disasters caused by storms. The US goverment spends approximately billions of dollars yearly to support and indemnify the victims of natural disasters. In 2011 and 2012 alone, damages from storms were evaluated at 188 billions dollars[@Weiss2013] and 1107 fatalities were officially recorded.

Every year, as it has been since 1826, NOAA(National Oceanographic Athmospheric Agency) collect and compiles data about storms-related occurences on the US soil. The data includes an estimation of property and crop damages, as well as a count of fatalities and injuries per event. All events are categorized by their type and additional location information is provided.

To determine what type of storm event has the greatest economical and public health impact, storm data from 1950 to November 2011 is obtained from NOAA, and Exploratory analysis is conducted on the data.

Data Processing

Data analysis and processing is conducted using the free and open source R statistical analysis toolset.The environment consist of RStudio Version 0.98.953 and the following R parameters:

```
R.Version()
```

```
## $platform
## [1] "x86_64-w64-mingw32"
##
## $arch
## [1] "x86_64"
##
## $os
## [1] "mingw32"
##
## $system
## [1] "x86_64, mingw32"
##
## $status
## [1] ""
##
## $major
## [1] "3"
##
## $minor
## [1] "1.1"
##
## $year
## [1] "2014"
##
## $month
## [1] "07"
##
## $day
## [1] "10"
##
## $`svn rev`
## [1] "66115"
##
## $language
## [1] "R"
##
## $version.string
## [1] "R version 3.1.1 (2014-07-10)"
##
## $nickname
## [1] "Sock it to Me"
```

To answer the two questions about the effect of storm events on public health and the economy, storm data from 1950 to November 2011 is obtained from here (https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2)

```
#  Link to the compressed data file on website
#fileURL <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
#download.file(fileURL, destfile="storm.bz2")
```

decompressed into the current folder,

```
#require(R.utils)
#bunzip2("./storm.bz2")
```

and the raw data is imported into the R environment

```
rawdata <- read.csv("repdata-data-StormData.csv")
```

A decription of imported data is available in the Storm Data Documentation (https://d396qusza40orc.cloudfront.net/repdata%2Fpeer2_doc%2Fpd01016005curr.pdf). A peek at the data structure should help identify all the attributes described in the document above.

```
str(rawdata)
```

```
## 'data.frame':    902297 obs. of  37 variables:
##  $ STATE__   : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ BGN_DATE  : Factor w/ 16335 levels "1/1/1966 0:00:00",..: 6523 6523 4242 11116 2224 2224
2260 383 3980 3980 ...
##  $ BGN_TIME  : Factor w/ 3608 levels "00:00:00 AM",..: 272 287 2705 1683 2584 3186 242 1683
3186 3186 ...
##  $ TIME_ZONE : Factor w/ 22 levels "ADT","AKS","AST",..: 7 7 7 7 7 7 7 7 7 7 ...
##  $ COUNTY    : num  97 3 57 89 43 77 9 123 125 57 ...
##  $ COUNTYNAME: Factor w/ 29601 levels "","5NM E OF MACKINAC BRIDGE TO PRESQUE ISLE LT MI",..
: 13513 1873 4598 10592 4372 10094 1973 23873 24418 4598 ...
##  $ STATE     : Factor w/ 72 levels "AK","AL","AM",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ EVTYPE    : Factor w/ 985 levels "   HIGH SURF ADVISORY",..: 834 834 834 834 834 834 834
834 834 834 ...
##  $ BGN_RANGE : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ BGN_AZI   : Factor w/ 35 levels "","  N"," NW",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ BGN_LOCATI: Factor w/ 54429 levels "","- 1 N Albion",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ END_DATE  : Factor w/ 6663 levels "","1/1/1993 0:00:00",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ END_TIME  : Factor w/ 3647 levels "","  0900CST",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ COUNTY_END: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ COUNTYENDN: logi  NA NA NA NA NA NA ...
##  $ END_RANGE : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ END_AZI   : Factor w/ 24 levels "","E","ENE","ESE",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ END_LOCATI: Factor w/ 34506 levels "","- .5 NNW",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ LENGTH    : num  14 2 0.1 0 0 1.5 1.5 0 3.3 2.3 ...
##  $ WIDTH     : num  100 150 123 100 150 177 33 33 100 100 ...
##  $ F         : int  3 2 2 2 2 2 2 1 3 3 ...
##  $ MAG       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ FATALITIES: num  0 0 0 0 0 0 0 0 1 0 ...
##  $ INJURIES  : num  15 0 2 2 2 6 1 0 14 0 ...
##  $ PROPDMG   : num  25 2.5 25 2.5 2.5 2.5 2.5 2.5 25 25 ...
##  $ PROPDMGEXP: Factor w/ 19 levels "","-","?","+",..: 17 17 17 17 17 17 17 17 17 17 ...
##  $ CROPDMG   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ CROPDMGEXP: Factor w/ 9 levels "","?","0","2",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ WFO       : Factor w/ 542 levels "","  CI","$AC",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ STATEOFFIC: Factor w/ 250 levels "","ALABAMA, Central",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ ZONENAMES : Factor w/ 25112 levels "","
                                                                                "| __truncated__,.
.: 1 1 1 1 1 1 1 1 1 1 ...
##  $ LATITUDE  : num  3040 3042 3340 3458 3412 ...
##  $ LONGITUDE : num  8812 8755 8742 8626 8642 ...
##  $ LATITUDE_E: num  3051 0 0 0 0 ...
##  $ LONGITUDE_: num  8806 0 0 0 0 ...
##  $ REMARKS   : Factor w/ 436774 levels "","-2 at Deer Park\n",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ REFNUM    : num  1 2 3 4 5 6 7 8 9 10 ...
```

Couple of inconsistencies ought to be noted here. The STATE attribute is being reported as a factor with 72 levels (denoting 72 states.) In addition, event types (EVTYPE) consist of 985 levels, which is wildly above the 40 types of events decribed in page 6 of the storm data documentation. These denotes for a data that is not quite tidy.

Before engaging in any data correction however, the raw data can be subsetted to contain only attributes that matter for upcomming analysis. Those attributes are identified below and can be aggregated by: * Event type and location + STATE + EVTYPE (Event Type) * Economic impacts indicators + PROPDMG(property damage),
+ PROPDMGEXP(Property damage magnitude) , + CROPDMG(crop Damage), + and CROPDMGEXP(crop damage magnitude), * Human health impact indicators + FATALITIES, + and INJURIES

```
rawData <- rawdata[, c("STATE","EVTYPE","PROPDMG","PROPDMGEXP","CROPDMG","CROPDMGEXP","FATALITI
ES", "INJURIES")]
```

Next, looking at unique values of STATE and EVTYPE in our new data frame, an observations is made that there are duplicates values that are differents only by the case of the letters. As such, factor levels “Hurricane” and “HURRICANE” are deemed to be differents factors though they indicate the same event. To solve this issue, all factors will be uppercased.

```
# transform factor to character, to uppercase,
# trim whitespaces, and convert back as a factor
rawData$STATE <- as.factor(toupper(str_trim(as.character(rawData$STATE))))
rawData$EVTYPE<- as.factor(toupper(str_trim(as.character(rawData$EVTYPE))))
```

Peeking into the rawData, we notice that though the number of STATE factors have not changed, EVTYPE factor has been reduced from 985 categories to 890. Further processing might be needed to reduce the number of factor to a reasonable number.
Let’s reconvert EVTYPE into a string,

```
rawData$EVTYPE<- as.character(rawData$EVTYPE)
```

extract and sorting event types as reported by rawData.

```
eventTypes<- unique(rawData$EVTYPE)
eventTypes<- eventTypes[order(eventTypes)]
rawData$EVTYPE<- as.factor(rawData$EVTYPE)
```

Peeking in the data again, we notice spelling errors. For example, items 10 and 11, item 12 and 13 describe respectively the same event types. Mispelled items would lead to additional unnecessary factors.

The magnitude fields PROPDMGEXP and CROPDMGEXP are reported in the data as characters

```
levels(rawData$PROPDMGEXP)
```

```
##  [1] ""  "-" "?" "+" "0" "1" "2" "3" "4" "5" "6" "7" "8" "B" "h" "H" "K"
## [18] "m" "M"
```

```
levels(rawData$CROPDMGEXP)
```

```
## [1] ""  "?" "0" "2" "B" "k" "K" "m" "M"
```

the letters represent exponents of the values. Here again, because of character cases, duplicate factors are included and must be corrected.

```
rawData$PROPDMGEXP <- as.factor(toupper(str_trim(as.character(rawData$PROPDMGEXP))))
rawData$CROPDMGEXP <- as.factor(toupper(str_trim(as.character(rawData$CROPDMGEXP))))
```

For example, K will denote an order of magnitude of 10^3, B for 10^6, 8 for 10^8. To determine the real value of the damage for processing purpose, a transformation is effected to create additional columns in the data: PROPDMGPOWERofTEN, CROPDMGPOWERofTEN, PROPDMGDOLLAR , CROPDMGDOLLAR, DAMAGEDOLLAR, TOTALFATALITIES

```
powerOfTen <- "'H'=100;'K'=1000;'M'=1000000;'B'=1000000000;'-'=0;'?'=0;'+'=0;'0'=1;'1'=10;'2'=1
00;'3'=1000;'4'=10000;'5'=100000;'6'=1000000;'7'=10000000;'8'=100000000"

rawData$PROPDMGPOWERofTEN <- as.numeric(recode(rawData$PROPDMGEXP,powerOfTen))
rawData$CROPDMGPOWERofTEN <- as.numeric(recode(rawData$CROPDMGEXP,powerOfTen))



rawData$PROPDMGDOLLAR <-  rawData$PROPDMG * rawData$PROPDMGPOWERofTEN
rawData$CROPDMGDOLLAR <-  rawData$CROPDMG * rawData$CROPDMGPOWERofTEN

rawData$DAMAGEDOLLAR <-  rawData$CROPDMGDOLLAR + rawData$PROPDMGDOLLAR
rawData$TOTALFATALITIES <-  rawData$FATALITIES + rawData$INJURIES
```

The data might require extensive cleaning and retransformation to be fully usable, however analysis will be performed on the data in its current status.

Results

To measure the effect of storms on public health, the total number of fatalities and injuries is calculated per storm type. The human cost is evaluated as follow:

```
HumanCost <- tapply(rawData$DAMAGEDOLLAR, rawData$EVTYPE, sum)
HumanCost <- HumanCost[order(HumanCost, decreasing = T)][1:5]
```

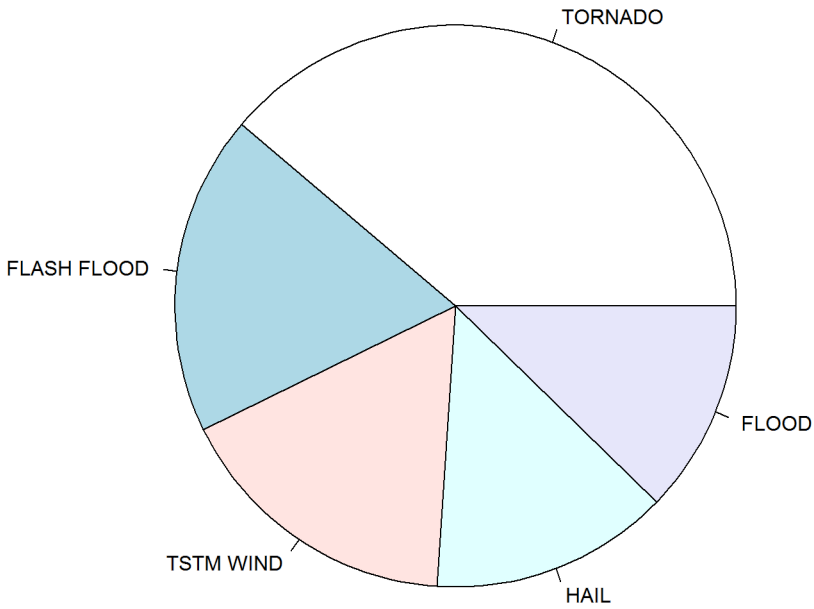
The economic effect of storms is measured by calculating the total amount of damage both to crops and properties as follows:

```
DamageCost <- tapply(rawData$CROPDMGDOLLAR, rawData$EVTYPE, sum)
DamageCost <- DamageCost[order(DamageCost, decreasing = T)][1:5]
```

and the results can be plotted.

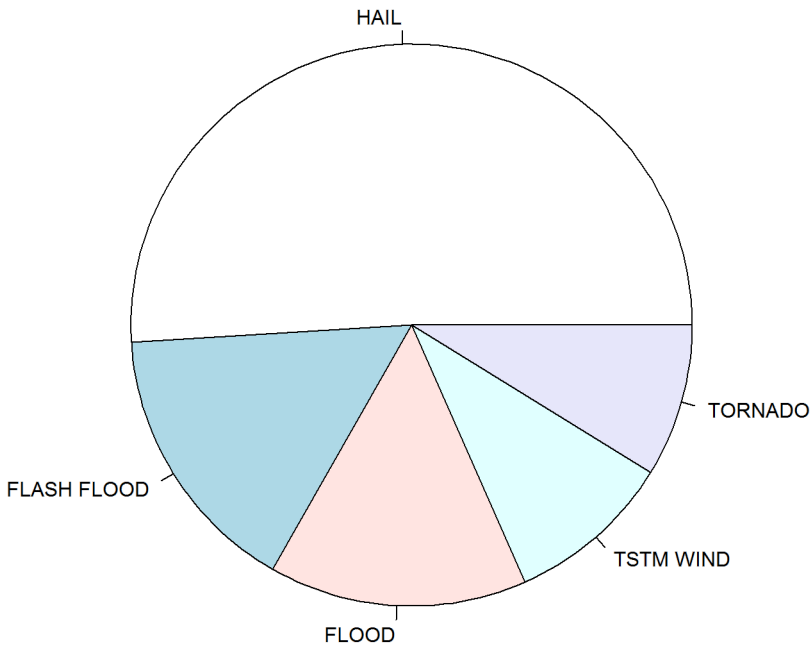
The firt plot is that of the five most important storm types and their effect on human cost.

```
pie(HumanCost)
```



The following plot is that of the effect of the five most important storm types on the economy

```
pie(DamageCost)
```



Conclusion

Accross the US, tornadoes have the greatest human impact with greater numbers of fatalities and injuries recorded. On the other hand, damages to crop and properties are overwhelmingly created by hail more than any other type of storm.

References