The Impact of Storm Events on Public Health and Damage – An Analysis of US Storm Data

Author: Jens Berkmann

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. In this report the U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database is explored and we found that tornados have the most harmful consequences to population health while flood and draught have the greatest economic consequences. In fact, for the whole time period of the database more than 5000 people died in the US as a results of a tornado and more than 80000 people were injured. Moreover, a total amount worth of roughly 100 Billion USD each of property and crop damage were caused by flood and draught, respectively.

Data Processing

Let us start by setting global options and loading the required R packages followed by dataset download.

```
#Global settings
library(knitr)
library(downloader)
library(plyr)
library(stringr)
library(ggplot2)
library(gridExtra)
```

```
## Loading required package: grid
```

```
knitr::opts_chunk$set(cache = TRUE)
knitr::opts_chunk$set(echo = TRUE)
# Environment for data analysis
sessionInfo()
```

```
## R version 3.1.1 (2014-07-10)
## Platform: i386-w64-mingw32/i386 (32-bit)
##
## locale:
## [1] LC_COLLATE=German_Germany.1252  LC_CTYPE=German_Germany.1252
## [3] LC_MONETARY=German_Germany.1252 LC_NUMERIC=C
## [5] LC_TIME=German_Germany.1252
##
## attached base packages:
## [1] grid      stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
## [1] gridExtra_0.9.1 ggplot2_1.0.0   stringr_0.6.2   plyr_1.8.1
## [5] downloader_0.3  knitr_1.6
```

```
##
## loaded via a namespace (and not attached):
##  [1] colorspace_1.2-4 digest_0.6.4     evaluate_0.5.5   formatR_0.10
##  [5] gtable_0.1.2     htmltools_0.2.4  MASS_7.3-33      munsell_0.4.2
##  [9] proto_0.3-10     Rcpp_0.11.2      reshape2_1.4     rmarkdown_0.2.49
## [13] scales_0.2.4     tools_3.1.1      yaml_2.1.13
```

```
URL <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
filename <- "tmp.csv.bz2"
download(URL, filename, mode="wb")
data <- read.csv("tmp.csv.bz2")
names(data)
```

```
##  [1] "STATE__"    "BGN_DATE"   "BGN_TIME"   "TIME_ZONE"  "COUNTY"
##  [6] "COUNTYNAME" "STATE"      "EVTYPE"     "BGN_RANGE"  "BGN_AZI"
## [11] "BGN_LOCATI" "END_DATE"   "END_TIME"   "COUNTY_END" "COUNTYENDN"
## [16] "END_RANGE"  "END_AZI"    "END_LOCATI" "LENGTH"     "WIDTH"
## [21] "F"          "MAG"        "FATALITIES" "INJURIES"   "PROPDMG"
## [26] "PROPDMGEXP" "CROPDMG"    "CROPDMGEXP" "WFO"        "STATEOFFIC"
## [31] "ZONENAMES"  "LATITUDE"   "LONGITUDE"  "LATITUDE_E" "LONGITUDE_"
## [36] "REMARKS"    "REFNUM"
```

From the descriptive information on the dataset available at https://d396qusza40orc.cloudfront.net/repdata%2Fpeer2_doc%2Fpd01016005curr.pdf we extract the relevant columns from the dataset that are required to answer the questions stated.

```
data <- data[,c("EVTYPE",   "FATALITIES", "INJURIES",
                            "PROPDMG", "PROPDMGEXP",
                            "CROPDMG", "CROPDMGEXP")]
```

According to the documentation the data in the columns *PROPDMGEXP* and *CROPDMGEXP* are supposed to contain the "exponents" K (Kilo), M (Million), and B (Billion) only to magnify the numbers in the columns *PROPDMG* and *CROPDMG*, respectively. However when inspecting the columns we find that other "exponential" values exist, however their frequency is not much relevant.

```
count(data$PROPDMGEXP)
```

```
##    x   freq
## 1     465934
## 2  -      1
## 3  ?      8
## 4  +      5
## 5  0    216
## 6  1     25
## 7  2     13
## 8  3      4
## 9  4      4
## 10 5     28
## 11 6      4
## 12 7      5
## 13 8      1
## 14 B     40
```

```
## 15 h      1
## 16 H      6
## 17 K 424665
## 18 m      7
## 19 M  11330
```

```
count(data$CROPDMGEXP)
```

```
##   x   freq
## 1   618413
## 2 ?      7
## 3 0     19
## 4 2      1
## 5 B      9
## 6 k     21
## 7 K 281832
## 8 m      1
## 9 M   1994
```

We decide to:

- convert k,m,b to capital letters and interpret them as K,M,B, i.e. treat them as exponents (exp) 10^{exp} with $exp = 3, 6, 9$
- treat numbers 0-9 as exponents 10^{exp}
- ignore the empty string "", as well as -","+","¿', i.e. treat them as an exponent $exp = 0$
- treat "h" and "H" as an exponent $exp = 0$ even though it could mean "hundreds" (but we do not know).

```
data$PROPDMGEXP <- toupper(data$PROPDMGEXP)
data$CROPDMGEXP <- toupper(data$CROPDMGEXP)
data$PROPDMGEXP <- str_trim(data$PROPDMGEXP)
data$CROPDMGEXP <- str_trim(data$CROPDMGEXP)
data$PROPDMGEXP <- gsub("[-+?hH]","0",data$PROPDMGEXP)
data$CROPDMGEXP <- gsub("[-+?hH]","0",data$CROPDMGEXP)
data[data$PROPDMGEXP=="",]$PROPDMGEXP <- "0"
data[data$CROPDMGEXP=="",]$CROPDMGEXP <- "0"
data$PROPDMGEXP <- gsub("K","3",data$PROPDMGEXP)
data$PROPDMGEXP <- gsub("M","6",data$PROPDMGEXP)
data$PROPDMGEXP <- gsub("B","9",data$PROPDMGEXP)
data$CROPDMGEXP <- gsub("K","3",data$CROPDMGEXP)
data$CROPDMGEXP <- gsub("M","6",data$CROPDMGEXP)
data$CROPDMGEXP <- gsub("B","9",data$CROPDMGEXP)
```

We then create 3 new columns which contain the costs in US$ for property damage and crop damage as well as the sum of those 2 values.

```
data$PROP<- data$PROPDMG  * 10^(as.numeric(data$PROPDMGEXP))
data$CROP<- data$CROPDMG  * 10^(as.numeric(data$CROPDMGEXP))
data$DMG <- data$PROP + data$CROP
```

The final pre-processed data can be summarized as follows:

```
data <- data[,c("EVTYPE",  "FATALITIES", "INJURIES",
                           "PROP", "CROP", "DMG")]
str(data)
```

```
## 'data.frame':    902297 obs. of  6 variables:
##  $ EVTYPE    : Factor w/ 985 levels "   HIGH SURF ADVISORY",..: 834 834 834 834 834 834 834 834 834 8
##  $ FATALITIES: num  0 0 0 0 0 0 0 0 1 0 ...
##  $ INJURIES  : num  15 0 2 2 2 6 1 0 14 0 ...
##  $ PROP      : num  25000 2500 25000 2500 2500 2500 2500 2500 25000 25000 ...
##  $ CROP      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ DMG       : num  25000 2500 25000 2500 2500 2500 2500 2500 25000 25000 ...
```

```
summary(data)
```

```
##                EVTYPE         FATALITIES     INJURIES
##  HAIL              :288661   Min.   :  0   Min.   :   0.0
##  TSTM WIND         :219940   1st Qu.:  0   1st Qu.:   0.0
##  THUNDERSTORM WIND: 82563   Median :  0   Median :   0.0
##  TORNADO           : 60652   Mean   :  0   Mean   :   0.2
##  FLASH FLOOD       : 54277   3rd Qu.:  0   3rd Qu.:   0.0
##  FLOOD             : 25326   Max.   :583   Max.   :1700.0
##  (Other)           :170878
##       PROP                CROP                DMG
##  Min.   :0.00e+00   Min.   :0.00e+00   Min.   :0.00e+00
##  1st Qu.:0.00e+00   1st Qu.:0.00e+00   1st Qu.:0.00e+00
##  Median :0.00e+00   Median :0.00e+00   Median :0.00e+00
##  Mean   :4.75e+05   Mean   :5.44e+04   Mean   :5.29e+05
##  3rd Qu.:5.00e+02   3rd Qu.:0.00e+00   3rd Qu.:1.00e+03
##  Max.   :1.15e+11   Max.   :5.00e+09   Max.   :1.15e+11
##
```

Processing of Data

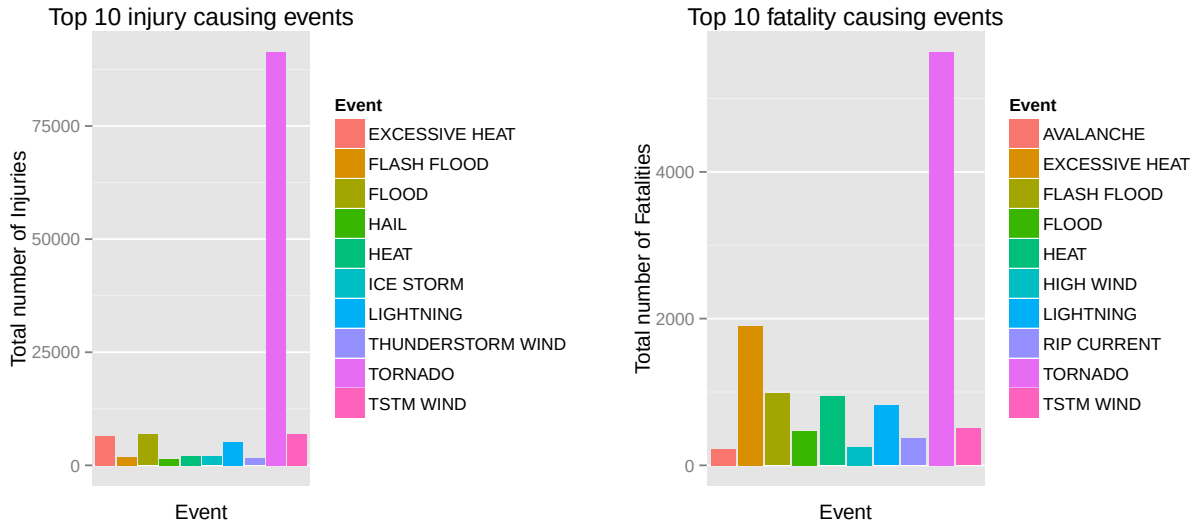
Let us now look at the sum of the individual variables “FATALITIES”, “INJURIES”, “PROP”, “CROP” depending on the type of event. We then rearrange the data in decreasing order from which we finally extract the top 10 rows for final plotting.

```
inju <- aggregate(data$INJURIES,   list(Event=data$EVTYPE), sum)
fata <- aggregate(data$FATALITIES, list(Event=data$EVTYPE), sum)
prop <- aggregate(data$PROP,       list(Event=data$EVTYPE), sum)
crop <- aggregate(data$CROP,       list(Event=data$EVTYPE), sum)
inju <- arrange(inju,desc(x))[1:10,]
fata <- arrange(fata,desc(x))[1:10,]
prop <- arrange(prop,desc(x))[1:10,]
crop <- arrange(crop,desc(x))[1:10,]
```

Results

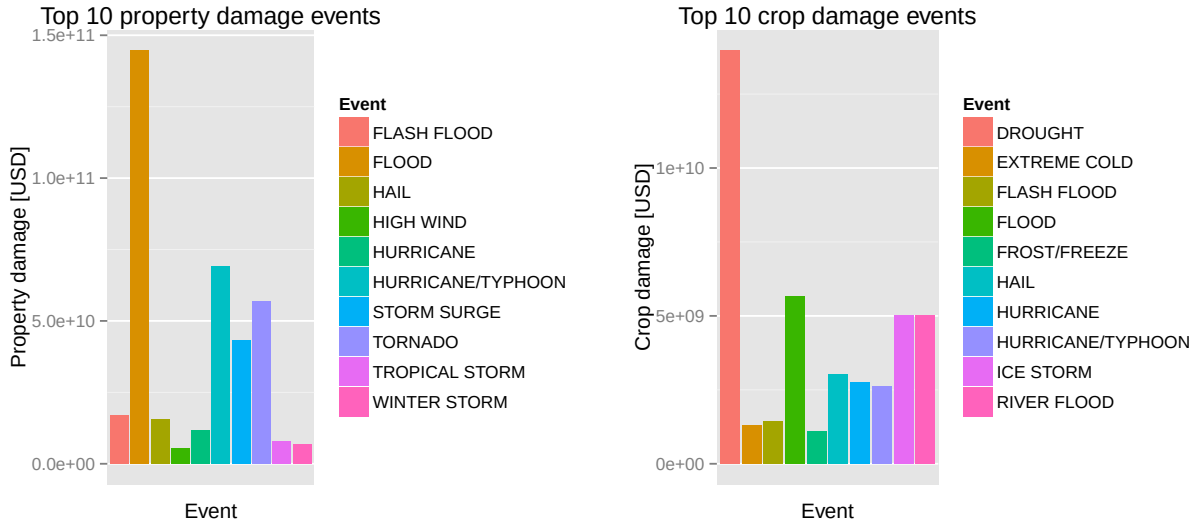
In the following figure the top 10 events are shown causing the highest total number of fatalities and injuries, respectively.

```
pic1 <- ggplot(inju, aes(x=Event, y=x, fill=Event)) +
     geom_bar(stat="identity") +
     scale_x_discrete(breaks=NULL) +
     labs(y="Total number of Injuries") +
     ggtitle("Top 10 injury causing events")
pic2 <- ggplot(fata, aes(x=Event, y=x, fill=Event)) +
     geom_bar(stat="identity") +
     scale_x_discrete(breaks=NULL) +
     labs(y="Total number of Fatalities") +
     ggtitle("Top 10 fatality causing events")
p<-grid.arrange(pic1, pic2, ncol=2)
```



In the final figure the top 10 events are shown causing the highest total number of damage of property or crop, respectively.

```
pic1 <- ggplot(prop, aes(x=Event, y=x, fill=Event)) +
     geom_bar(stat="identity") +
     scale_x_discrete(breaks=NULL) +
     labs(y="Property damage [USD]") +
     ggtitle("Top 10 property damage events")
pic2 <- ggplot(crop, aes(x=Event, y=x, fill=Event)) +
     geom_bar(stat="identity") +
     scale_x_discrete(breaks=NULL) +
     labs(y="Crop damage [USD]") +
     ggtitle("Top 10 crop damage events")
p<-grid.arrange(pic1, pic2, ncol=2)
```



It is observed that tornados are most harmful for population health. Flood causes greatest economical consequences for property damage in the US while the highest costs for crop damage are caused by draught.