Title: Since 1950, three types of bad weather have caused the bulk of damage and injury in the US

Synopsis: summary of analysis and findings

Of the most damaging forms of weather in the US, the three forms of weather that are responsible for the most injuries and death are, in order, Tornados, Thunderstorms, and Excessive Heat. The weather which causes the most property and crop damage is Thunderstorms, Tornados, and Flooding. The data was a little difficult to combine, because over the course of collection and labeling some variance in names and categorization occured over the years during data capture and entry. Also, advances in weather tracking and recording have made more recent data more accurate in classification than historically possible. The data also does not take into account the inflation of the USD, which is the monetary basis of the damage cost, so this means that $10,000 damage in 1951 is not adjusted for actual value versus an equivolent of $100,000 damage in 2001.

Data download and loading

Download and load massive dataset containing all the storm data since 1950

```
temp <- tempfile()

download.file("https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2", method = "curl", temp)

data <- read.csv(bzfile(temp))

unlink(temp)
```

Data Processing

Many Event Types are the same but written differently. It would be too burdensome to audit every possible similar event, so instead we have iterated through the report many times, and in each iteration, evaluated the duplicates that appeared in our top 20 list and cleaned them up until no duplicates were found. It was decided that this was “close enough” to be “good enough” for the level of science we are trying to achieve.

```
library(plyr)

data$EVTYPE <- as.factor( mapvalues( data$EVTYPE, from= c(" TSTM WIND"," LIGHTNING","LIGHTNING", "THUNDERSTORM WIND","THUNDERSTORM WINDS"), to= rep("THUNDERSTORM", 5) ) )
```

```r
data$EVTYPE <- as.factor( mapvalues( data$EVTYPE, from= c("EXCESSIVE HEAT","HEAT","HEAT WAVE", " HEAT"), to= rep("EXCESSIVE HEAT",4) ) )
```

```
## The following `from` values were not present in `x`:  HEAT
```

```r
data$EVTYPE <- as.factor( mapvalues( data$EVTYPE, from= c("FLOOD", "FLASH FLOOD", " FLASH FLOOD", "FLASH FLOODING", "URBAN/SML STREAM FLD", "FLOOD/FLASH FLOOD", "RIVER FLOOD", "URBAN FLOOD", " COASTAL FLOOD"), to= rep("FLOOD/FLASH FLOOD",9) ) )

data$EVTYPE <- as.factor( mapvalues( data$EVTYPE, from= c("ICE STORM", "WINTER STORM", "HEAVY SNOW", "BLIZZARD", "WINTER WEATHER"), to= rep("WINTER STORM",5) ) )

data$EVTYPE <- as.factor( mapvalues( data$EVTYPE, from= c("WILDFIRE", "WILD/FOREST FIRE"), to= rep("WILD/FOREST FIRE",2) ) )

data$EVTYPE <- as.factor( mapvalues( data$EVTYPE, from= c("HIGH WIND", "HIGH WINDS", "STRONG WIND", "STRONG WINDS", " WIND"), to= rep("HIGH WIND",5) ) )
```

Filter down all the columns and sum both fatalities and injuries by event type, then remove all event types that had no injuries or fatalities.

```r
sdata <-aggregate(x=data[, c("FATALITIES","INJURIES")], by = list(data$EVTYPE), FUN = sum)

sdata$total <- sdata$FATALITIES + sdata$INJURIES

sdata <- sdata[ sdata$total > 0, ]

names(sdata)[1] <- "Event Type"

str(sdata)
```

```
## 'data.frame':    216 obs. of  4 variables:

##  $ Event Type: Factor w/ 985 levels "   HIGH SURF ADVISORY",..: 18 19 29 30 42 44 49 54 56 57 ...

##  $ FATALITIES: num  1 224 1 101 1 1 0 3 2 1 ...

##  $ INJURIES  : num  0 170 24 805 1 13 2 2 0 0 ...

##  $ total     : num  1 394 25 906 2 14 2 5 2 1 ...
```

Filter down all the columns and create a new dataset that only contains the event type, and damage amounts. Then convert the notated damage amounts to numeric and sum total cost of damage to property and agriculture.

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

levels(mdata$PROPDMGEXP) <- c(levels(mdata$PROPDMGEXP), "1000", "1000000", "1000000000")

mdata$PROPDMGEXP[mdata$PROPDMGEXP == "K" || mdata$PROPDMGEXP == "k"] <- "1000"

mdata$PROPDMGEXP[mdata$PROPDMGEXP == "M" || mdata$PROPDMGEXP == "m"] <- "1000000"
```

```
mdata$PROPDMGEXP[mdata$PROPDMGEXP == "B" || mdata$PROPDMGEXP == "b"] <- "1000000000"

levels(mdata$CROPDMGEXP) <- c(levels(mdata$CROPDMGEXP), "1000", "1000000", "1000000000")

mdata$CROPDMGEXP[mdata$CROPDMGEXP == "K" || mdata$CROPDMGEXP == "k"] <- "1000"

mdata$CROPDMGEXP[mdata$CROPDMGEXP == "M" || mdata$CROPDMGEXP == "m"] <- "1000000"

mdata$CROPDMGEXP[mdata$CROPDMGEXP == "B" || mdata$CROPDMGEXP == "b"] <- "1000000000"


#filter out rows that do not have an expression for either property or crop damages

mdata <- mdata[mdata$PROPDMGEXP %in% c("1000", "1000000", "1000000000") | mdata$CROPDMGEXP %in% c("1000", "1000000", "1000000000"), ]


#Calculate subtotals, and then total of economic impact

mdata$PROPTOTAL <- mdata$PROPDMG * as.numeric(mdata$PROPDMGEXP)

mdata$CROPTOTAL <- mdata$CROPDMG * as.numeric(mdata$CROPDMGEXP)

mdata <- aggregate(x=mdata[, c("PROPTOTAL","CROPTOTAL")], by = list(data$EVTYPE), FUN = sum)

mdata$total <- mdata$PROPTOTAL + mdata$CROPTOTAL

mdata <- mdata[ mdata$total > 0, ]

names(mdata)[1] <- "Event Type"
```

Results

Across the US the top 20 most harmful events, by population health are listed here:

```
ord <- order(sdata$total, decreasing=TRUE)

head( sdata[ord, ] ,n=20)
```

```
##            Event Type FATALITIES INJURIES total

## 831           TORNADO       5633    91346 96979

## 752      THUNDERSTORM       1518    14595 16113

## 130    EXCESSIVE HEAT       1903     6525  8428

## 170             FLOOD        470     6789  7259

## 275              HEAT        937     2100  3037

## 153       FLASH FLOOD        978     1777  2755
```

```
## 427          ICE STORM         89      1975  2064
## 968       WINTER STORM        206      1321  1527
## 359          HIGH WIND        248      1137  1385
## 244               HAIL         15      1361  1376
## 411  HURRICANE/TYPHOON         64      1275  1339
## 310         HEAVY SNOW        127      1021  1148
## 953           WILDFIRE         75       911   986
## 30            BLIZZARD        101       805   906
## 188                FOG         62       734   796
## 584        RIP CURRENT        368       232   600
## 951   WILD/FOREST FIRE         12       545   557
## 585       RIP CURRENTS        204       297   501
## 278          HEAT WAVE        172       309   481
## 117         DUST STORM         22       440   462
```

Across the US the top 20 economic consequences and their cost are listed here:

```
ord <- order(mdata$total, decreasing=TRUE)
head( mdata[ord, ] ,n=20)
```

```
##            Event Type PROPTOTAL CROPTOTAL    total
## 457      THUNDERSTORM  65267506   1389403 66656909
## 824           TORNADO  64245163    700120 64945284
## 153 FLOOD/FLASH FLOOD  47875308   2509796 50385104
## 241              HAIL  13773868   4061557 17835424
## 30       WINTER STORM   6925835     42610  6968444
## 353         HIGH WIND   6494631    122226  6616857
## 943  WILD/FOREST FIRE   2476086     60664  2536750
## 669       STRONG WIND   1259876     11445  1271321
## 370        HIGH WINDS   1112500     12395  1124895
## 285        HEAVY RAIN   1016843     79306  1096148
## 838    TROPICAL STORM    968474     42640  1011114
## 663       STORM SURGE    387870        35   387905
## 435         LANDSLIDE    379239       299   379538
```

```
## 396          HURRICANE    310274     42854    353127
## 95             DROUGHT     81981    262190    344171
## 583        RIVER FLOOD    277114     24472    301586
## 433  LAKE-EFFECT SNOW    282820         0    282820
## 894        URBAN FLOOD    265860      6464    272324
## 54       COASTAL FLOOD    252217         0    252217
## 964     WINTER WEATHER    239498       135    239633
```

Study about most influence weather events in EEUU

This report aims to estimate which of the meteorological events is the most dangerous to human health and which of these events is the more economic damage caused in society. This work belongs to the module “reproducible research” Hopkins University in the Peer 2 Assignment. Data are provided by the database tormetas https://d396qusza40orc.cloudfront.net/repdata% 2Fdata% 2FStormData.csv.bz2. In the data processing has tried to limit the time to have a comparable sample, then select the attributes for the accomplishment of work, study the economic damage and get data that respond to events more meterologicos affect humans in the USA. The study provides evidence that tornadoes are generally the most dangerous events

First we load the data to analysis:

```
library(ggplot2)
a <-
read.csv("/home//ines.huertas/Escritorio/ADA/data_scientist/curso5/program
_assigment2/repdata-data-StormData.csv")
library(ggplot2)
```

We create the new dataset with the values

that we will use as we will base our study on the economic impact and the health of the people we have decided to use for the study variables related to it. Variables choosen: EVTYPE INJURIES FATALITIES PROPDMG PROPDMGEXP CROPDMG CROPDMGEXP

```
a <- subset(storms, select = c("EVTYPE", "FATALITIES", "INJURIES",
"PROPDMG",
    "PROPDMGEXP", "CROPDMG", "CROPDMGEXP"))
## Error: objeto 'storms' no encontrado
```

In the case of fields FATALITIES and INJURIES and does not require any transformation of the variables in terms of the fields of economic description if it is necessary for a treat as the variables appear ladop and secondly the exponent to which such amounts quantity, for it will create a new variable that is in a unique field that quantity. The exponents appear as factor type, we create a function to transform it into its economic value.

```
unique(a$CROPDMGEXP)
## [1]   M K m B ? 0 k 2
## Levels:  ? 0 2 B k K m M
unique(a$PROPDMGEXP)
##  [1] K M   B m + 0 5 6 ? 4 2 3 h 7 H - 1 8
## Levels:  - ? + 0 1 2 3 4 5 6 7 8 B h H K m M
```

Alphabetical characters used to signify magnitude include “K” for thousands, “M” for millions, and “B” for billions, there is no reference for others letters, we use unit “1” for rest.

```
# Function replace exponente
exponent <- function(expo) {
    result <- 1
    result = switch(expo, K = 1000, k = 1000, M = 1e+06, m = 1e+06, B =
1e+09,
        b = 1e+09, 1)
    print(result)

}

# expon<-sapply(a$CROPDMGEXP, exponent) VALCROPD<-a$CROPDMG*expon
# expon<-sapply(a$PROPDMG, exponent) VALPROPD<-a$PROPDMG*expon Add this
new
# values to subset a<-cbind(VALCROPD,a) a<-cbind(VALPROPD,a)
```

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

En el codebook encontramos el campo EVTYPE que indica el tipo de evento, junto con el campo FATALITIES e INJURIES, utilizaremos estos dos campos para valorar aquellos eventos mas harmfull

We obtain the number of deaths by type of event to see which are the most aggressive

```
resFatalities <- aggregate(a$INJURIES, list(a$EVTYPE), sum)
ord.NumFac <- order(resFatalities$x, decreasing = TRUE)
TopFac <- resFatalities[ord.NumFac, ]
```

We obtain the number of injuries by event type to see which are the most affected

```
resInj <- aggregate(a$INJURIES, list(a$EVTYPE), sum)
ord.NumInj <- order(resInj, decreasing = TRUE)
TopInj <- resFatalities[ord.NumInj, ]
```

Plot the graphics:

```
ggplot(a, aes(x = EVTYPE, y = FATALITIES)) + geom_point(colour = "red",
fill = "#FFCC66") +
    labs(title = "Accumulated deaths due to severe weather events in USA")
+
    theme(axis.text.x = element_text(angle = 90, vjust = 0.5, size = 7))
```

```
ggplot(a, aes(x = EVTYPE, y = INJURIES)) + geom_point(colour = "red", fill
= "#FFCC66") +
    labs(x = "Event") + labs(y = "Number of Injuries") + labs(title =
"Accumulated injuries due to severe weather events in USA") +
    theme(axis.text.x = element_text(angle = 90, vjust = 0.5, size = 7))
```

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

In this case we will operate as above but use the fields that represent an economic index PROPDMG, PROPDMGEXP, CROPDMG, CROPDMGEXP First procesad data, the field PROPDMGEXP And CROPPDMGEXP are indicators of unit

RepResearch_Peer2.Rmd

Libardo Lopez

Friday, July 25, 2014

Reproducible Research: Peer Assessment 2

Impact of weather events on public health and economics in USA

Synopsis:

Storms and other severe weather events can cause both public health and economic problems for communities. Many severe events can result in fatalities, injuries, and property damage, and preventing such outcomes to the extent possible is a key concern.
This project involves exploring the U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database.

The events in the database start in the year 1950 and end in November 2011; but the accuracy and completness of some of the older data is questionable.
With this in mind, lets restrict the analysis since 1980.

The original source of this data is:

- Dataset: Storm data [46.8MB]

The dataset is stored in a comma-separated-value (CSV) file and contains 902,297 observations with 37 variables.

- National Weather Service Storm Data Documentation Documentation
- National Climatic Data Center Storm Events FAQ FAQ

```
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 3.1.1
```

```
opts_knit$set(fig.keep='high', fig.path='figures/', dev='png', fig.width = 9, fig.height = 5, warning=FALSE, message=FALSE)
```

Setting, Loading and Transforming data

NOTE: Be sure to have the zip folder in the same working directory; in my case "G:/Proyectos/2014/Libardo/Peer2/"; please adjust it as your needs.

Load the dataset from the zip file and convert the string dates to R date-time format. Also Preselect variables with info.

```
Sys.setlocale("LC_TIME", "C") #change my local time to english
```

```
## [1] "C"
```

```
setwd("G:/Proyectos/2014/Libardo/Peer2/")

library(data.table)
```

```
## Warning: package 'data.table' was built under R version 3.1.1
```

```
file <- bzfile('repdata-data-StormData.csv.bz2')

data <- data.table(read.csv(file, stringsAsFactors=FALSE), as.is = TRUE, na.strings
= "")

data$EVTYPE <- as.factor(data$EVTYPE)

dim(data)
```

```
## [1] 902297     39
```

Preselect variables with relevant info

```
data <- data[INJURIES > 0 | FATALITIES > 0 | PROPDMG > 0 | CROPDMG > 0,

         list(COUNTY, COUNTYNAME, STATE, BGN_DATE,

              EVTYPE, FATALITIES, INJURIES, PROPDMG, PROPDMGEXP, CROPDMG, CROPDMGEX
P, LONGITUDE, LATITUDE )]

dim(data)
```

```
## [1] 254633     13
```

This dataset goes back to 1950, but the accuracy and completness of some of the older data is questionable. With this in mind, lets restrict my analysis since 1980.

format BGN_DATE to date.

```
library(stringr)
```

```
## Warning: package 'stringr' was built under R version 3.1.1
```

```
dates <- str_extract(data[["BGN_DATE"]], "\\d+/\\d+/\\d+")

dates <- as.Date(dates, format="%m/%d/%Y")

data <- cbind(data, dates)



cutoff <- as.Date("01/01/1980", format="%m/%d/%Y")

data <- data[dates >= cutoff, ]

dim(data)
```

```
## [1] 237782     14
```

Tansform variables to monetary value.

```
dt <- data[, list(PROPDMG = ifelse(PROPDMGEXP == "B", 1e+09 * PROPDMG,
                                   ifelse(PROPDMGEXP == "M", 1e+06 * PROPDMG,
                                          ifelse(PROPDMGEXP == "K", 1000 *PROPDMG,
0))),
                  CROPDMG = ifelse(CROPDMGEXP == "B", 1e+09 * CROPDMG,
                                   ifelse(CROPDMGEXP == "M", 1e+06 * CROPDMG,
                                          ifelse(CROPDMGEXP == "K", 1000 *CROPDMG,
0))),  INJURIES, FATALITIES, EVTYPE, COUNTY, COUNTYNAME, STATE, dates, LONGITUDE, L
ATITUDE)]


dt <- dt[INJURIES > 0 | FATALITIES > 0 | PROPDMG > 0 | CROPDMG > 0,
         list(COUNTY, COUNTYNAME, STATE, dates,
              EVTYPE, FATALITIES, INJURIES, PROPDMG, CROPDMG, LONGITUDE, LATITUDE)]
rm(data)
dim(dt)
```

```
## [1] 237468     11
```

Results

Summaries

By Event Type, summarize the total results since 1980 to nov 2011.

```
event_summ_by_EV = dt[
        ,
        list(
                Injur_EV=sum(INJURIES),
                Fatal_EV=sum(FATALITIES),
                Econ_cost_EV=sum(PROPDMG, CROPDMG)),
        by="EVTYPE"
        ]
event_summ_by_EV <- event_summ_by_EV[order(-Fatal_EV, -Injur_EV, -Econ_cost_EV, EVT
YPE )]
```

```
head(event_summ_by_EV,10)
```

```
##              EVTYPE Injur_EV Fatal_EV Econ_cost_EV
##  1:         TORNADO    37971     2274    4.404e+10
##  2: EXCESSIVE HEAT      6525     1903    5.002e+08
##  3:     FLASH FLOOD      1777      978    1.756e+10
##  4:            HEAT      2100      937    4.033e+08
##  5:       LIGHTNING      5230      816    9.408e+08
##  6:       TSTM WIND      6957      504    5.039e+09
##  7:           FLOOD      6789      470    1.503e+11
##  8:     RIP CURRENT       232      368    1.000e+03
##  9:       HIGH WIND      1137      248    5.909e+09
## 10:       AVALANCHE       170      224    3.722e+06
```

By State, summarize the total results since 1980 to nov 2011.

```
event_summ_by_State = dt[
        ,
        list(
                Injur_St=sum(INJURIES),
                Fatal_St=sum(FATALITIES),
                Econ_cost_St=sum(PROPDMG, CROPDMG)),
        by="STATE"
        ]
event_summ_by_State <- event_summ_by_State[order(-Fatal_St, -Injur_St, -Econ_cost_S
t, STATE )]
head(event_summ_by_State,10)
```

```
##     STATE Injur_St Fatal_St Econ_cost_St
##  1:    IL     2776     1287    1.364e+10
##  2:    TX    11744      976    3.289e+10
##  3:    PA     2949      838    5.350e+09
##  4:    FL     3918      692    4.516e+10
##  5:    MO     7165      616    7.471e+09
##  6:    AL     5580      573    1.729e+10
```

```
## 7:     CA     3251      550    1.271e+11
## 8:     NC     2916      376    1.021e+10
## 9:     TN     2964      360    6.464e+09
## 10:    NY     1298      340    4.925e+09
```

By State and Type of Event, summarize the total results since 1980 to nov 2011.

```
event_by_State_EVTYPE = dt[ ,
                    list(Injur_St_Ev=sum(INJURIES),
                         Fatal_St_Ev=sum(FATALITIES),
                         Econ_cost_St_Ev=sum(PROPDMG, CROPDMG)),
                         by=list(STATE, EVTYPE)]
event_by_State_EVTYPE <- event_by_State_EVTYPE[order(-Fatal_St_Ev, -Injur_St_Ev, -Econ_cost_St_Ev, EVTYPE, STATE )]
head(event_by_State_EVTYPE,10)
```

```
##     STATE         EVTYPE Injur_St_Ev Fatal_St_Ev Econ_cost_St_Ev
##  1:    IL           HEAT         241         653       4.650e+05
##  2:    AL        TORNADO        4767         406       5.852e+09
##  3:    PA EXCESSIVE HEAT         320         359       0.000e+00
##  4:    IL EXCESSIVE HEAT         352         330       0.000e+00
##  5:    TX EXCESSIVE HEAT          13         269       2.000e+05
##  6:    MO        TORNADO        2497         250       4.362e+09
##  7:    TN        TORNADO        2510         207       1.426e+09
##  8:    MO EXCESSIVE HEAT        3525         190       3.790e+05
##  9:    TX    FLASH FLOOD         587         177       9.678e+08
## 10:    FL    RIP CURRENT         149         172       0.000e+00
```

1 Across the United States, which types of events are most harmful with respect to population health?
Answer:
For public health, tornado was the most harmful event with 2274 fatalities, 37911 injuries and an economic cost in excess of $4.404 e 10, during the last 31 years (1980 to 2011).

2 Across the United States, which types of events have the greatest economic consequences?
Answer:
For economy, flood events have the greatest impact, with 470 fatalities, 6789 injuries and an economic cost in excess of $1.503 e 11, during the last 31 years (1980 to 2011).

Aditional findings

By State

IL is the most impacted by events, with 1287 fatalities, 2776 injuries and a cost about $1.346 e 10.
CA is the most expensive with 550 fatalities, 3251 injuries anda a cost about $1.271 e 11.

Plots

```
econ <- dt[, list(econ = sum(PROPDMG, CROPDMG, na.rm = TRUE)), by = EVTYPE][order(-econ)][1:10]

par(mfrow = c(1, 1), cex.axis = 0.7, cex.main = 1, mar = c(10, 4, 2, 1),

    oma = c(1, 1, 1, 1))

barplot(econ$econ/1e+06, names.arg = econ$EVTYPE, las = 2, col = "blue", main = "Economic Costs (in million) by Event type")
```

```
rm(econ)
```