John Hopkins University – Data Science Specialization – Reprodicible Research Course – Solution to Project 2

Dr. Guy Cohen

July 25, 2015

U.S. Weather Event Data (1950-2011) - Damage to Population Health and Ecomomic Consequences

Synopsis

In this document, we analyze the U.S. National Oceanic and Atmospheric Administration’s (NOAA) storm database. We look for event types that inflicted the most harm to population health. We also look for event types that caused the most financial damage. We plot our results in several barplots. We see that tornadoes produce the greatest average number of fatalities per event, 25. Tornadoes are followed by cold and snow, which produce 14 fatalities on average. The event type that produces the greatest number of injuries per event is “tropical storm Gordon” with 43 injuries per event on average, closely followed by wild fires with 37.5 injuries on average. With respect to financial damage, we find that tornadoes produce the greatest average property damage per event, $1.6B. The greatest crop damage per event is caused by excessive wetness. This damage is $142M.

Data Processing

First, we find the classes of the features in the database by reading the first 10,000 lines.

```r
invisible(memory.limit(size = 4000))
data <- read.table("repdata_data_StormData.csv.bz2", header = TRUE, sep=",",
                   quote = "\"", na.strings = "?", nrows = 10000,
                   stringsAsFactors = FALSE)
classes <- sapply(data, class)
classes[classes == "logical"] <- "character"
```

Second, we make a few manual corrections to the classes and load the entire database to memory.

```r
classes[c("BGN_DATE", "BGN_TIME", "END_DATE", "END_TIME", "F")] <- "character"
data <- read.table("repdata_data_StormData.csv.bz2", header = TRUE, sep=",",
                   quote = "\"", na.strings = "?", colClasses = classes,
                   stringsAsFactors = FALSE)
```

Third, we convert date columns to “Date”" class and convert relevant “numeric” class columns to “integer” class.

```r
data$BGN_DATE <- as.Date(data$BGN_DATE, "%m/%d/%Y")
data$END_DATE <- as.Date(data$END_DATE, "%m/%d/%Y")
integerClasses <- sapply(names(classes[classes == "numeric"]), function(x) {
    ifelse(all(data[x] - floor(data[x]) == 0.0, na.rm = TRUE), x, NA)})
integerClasses <- c(integerClasses[!is.na(integerClasses)], "F")
data[integerClasses] <- sapply(data[integerClasses], as.integer)
```

Fourth, we combine the “PROPDMG” (property damage) and “PROPDMGEXP” (property damage exponent) features, as well as the “CROPDMG” (crop damage) and “CROPDMGEXP” (crop damage exponent) features to two single features, “PROPDMG” and “CROPDMG” in units of US dollars. The “National Weather Service Instruction 10-1605”, which contains the codebook for this data base, states on page 12: “Alphabetical characters used to signify magnitude include”K" for thousands, “M” for millions, and “B” for billions." Thus, we denote any value other than “”, “K”, “M” or “B” as a missing value.

```
data$PROPDMGEXP <- toupper(data$PROPDMGEXP)
data$PROPDMGEXP[!(data$PROPDMGEXP %in% c("K", "M", "B", ""))] <- NA
data$PROPDMGEXP[data$PROPDMGEXP == ""] <- "0"
data$PROPDMGEXP[data$PROPDMGEXP == "K"] <- "1e3"
data$PROPDMGEXP[data$PROPDMGEXP == "M"] <- "1e6"
data$PROPDMGEXP[data$PROPDMGEXP == "B"] <- "1e9"

data$CROPDMGEXP <- toupper(data$CROPDMGEXP)
data$CROPDMGEXP[!(data$CROPDMGEXP %in% c("K", "M", "B", ""))] <- NA
data$CROPDMGEXP[data$CROPDMGEXP == ""] <- "0"
data$CROPDMGEXP[data$CROPDMGEXP == "K"] <- "1e3"
data$CROPDMGEXP[data$CROPDMGEXP == "M"] <- "1e6"
data$CROPDMGEXP[data$CROPDMGEXP == "B"] <- "1e9"

data$PROPDMG <- data$PROPDMG * as.numeric(data$PROPDMGEXP)
data$CROPDMG <- data$CROPDMG * as.numeric(data$CROPDMGEXP)
```

Fifth, we change the event type column (“EVTYPE”) to uppercase, since we later use it as a factor.

```
data$EVTYPE <- toupper(data$EVTYPE)
```

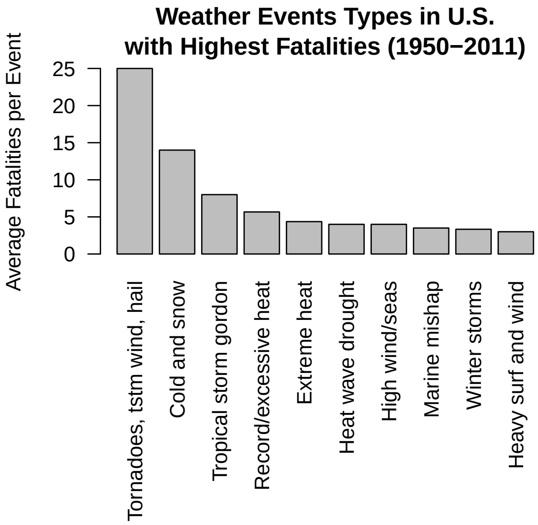
Results

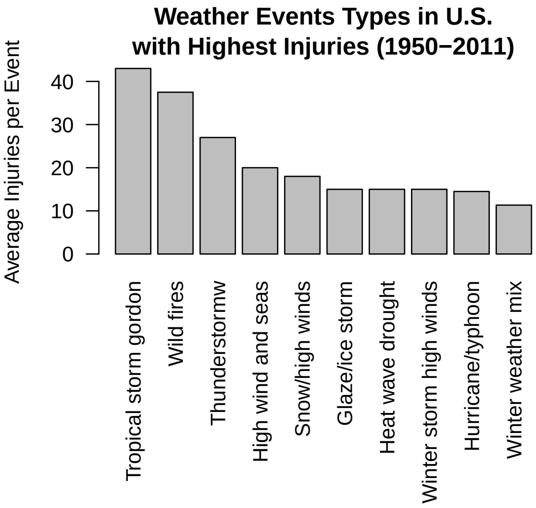
Events Types Across the United States That Are Most Harmful to Population Health

We examine the event types and look for the events most harmful with respect to the population health. Since people can, in most cases, heal from injuries, the number of fatalities is a better measure of damage to population health. In the barplots below, we present the 10 event types with the highest average number of fatalities per event and the 10 event types with the highest average number of injuries per event.

```
capitalize <- function(x) {
    paste(toupper(substr(x, 1, 1)), tolower(substr(x, 2, nchar(x))), sep="")
}
data[c("FATALITIES","INJURIES")] <-
    sapply(data[c("FATALITIES","INJURIES")], as.numeric)
aveFatalities <- aggregate(data = data, FATALITIES ~ EVTYPE, FUN = mean, na.rm = TRUE)
aveFatalities <- aveFatalities[order(aveFatalities$FATALITIES, decreasing = TRUE), ]
aveFatalities$EVTYPE <- sapply(aveFatalities$EVTYPE, capitalize)
aveInjuries <- aggregate(data = data, INJURIES ~ EVTYPE, FUN = mean, na.rm = TRUE)
aveInjuries <- aveInjuries[order(aveInjuries$INJURIES, decreasing = TRUE), ]
aveInjuries$EVTYPE <- sapply(aveInjuries$EVTYPE, capitalize)
par(las = 2, mfrow=c(1, 2), mar=c(12, 4, 3, 3))
with(aveFatalities[1:10, ], barplot(FATALITIES, names.arg = EVTYPE,
                                    ylab = "Average Fatalities per Event"))
title(main = "Weather Events Types in U.S.\nwith Highest Fatalities (1950-2011)")
```

```
with(aveInjuries[1:10, ], barplot(INJURIES, names.arg = EVTYPE,
                                  ylab = "Average Injuries per Event"))
title(main = "Weather Events Types in U.S.\nwith Highest Injuries (1950-2011)")
```

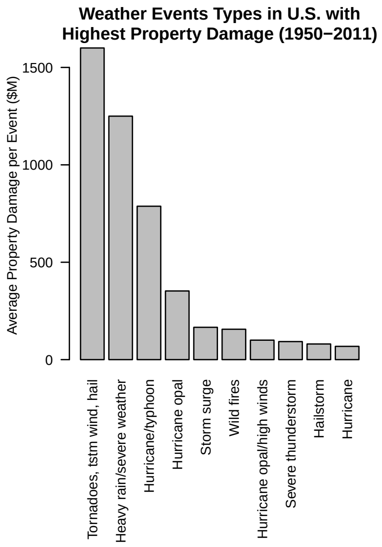




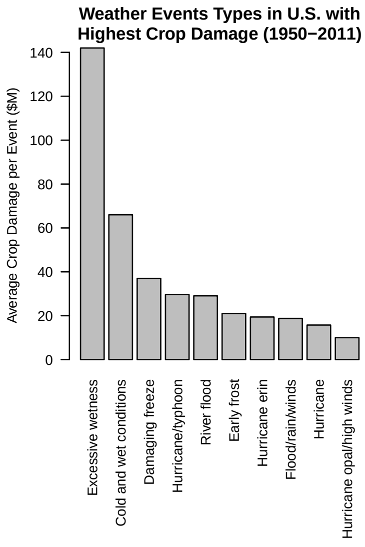
Events Types Across the United States with the Greatest Economic Consequences

We examine the event types and look for the events with the greatest economic consequences. Two measures of financial damage are given in the database, property damage ("PROPDMG") and crop damage ("CROPDMG"). In the barplots below, we present the 10 event types with the highest average property damage per event, the 10 event types with the highest crop damage per event and the 10 event types with the highest total (property plus crop) damage per event.

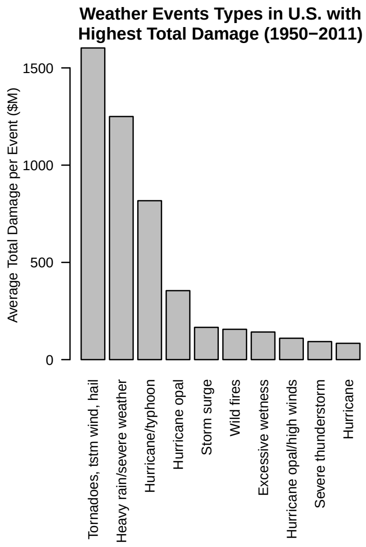
```
data$TOTDMG <- data$PROPDMG + data$CROPDMG
aveDamage <- aggregate(data = data, cbind(PROPDMG,CROPDMG,TOTDMG) ~ EVTYPE,
                       FUN = mean, na.rm = TRUE)
aveDamage$EVTYPE <- sapply(aveDamage$EVTYPE, capitalize)

par(las = 2, mfrow=c(1, 3), mar=c(12, 4, 3, 3))
aveDamage <- aveDamage[order(aveDamage$PROPDMG, decreasing = TRUE), ]
with(aveDamage[1:10, ], barplot(PROPDMG/10^6, names.arg = EVTYPE,
                                ylab = "Average Property Damage per Event ($M)"))
title(main = "Weather Events Types in U.S. with\nHighest Property Damage (1950-2011)")

aveDamage <- aveDamage[order(aveDamage$CROPDMG, decreasing = TRUE), ]
with(aveDamage[1:10, ], barplot(CROPDMG/10^6, names.arg = EVTYPE,
                                ylab = "Average Crop Damage per Event ($M)"))
title(main = "Weather Events Types in U.S. with\nHighest Crop Damage (1950-2011)")

aveDamage <- aveDamage[order(aveDamage$TOTDMG, decreasing = TRUE), ]
with(aveDamage[1:10, ], barplot(TOTDMG/10^6, names.arg = EVTYPE,
                                ylab = "Average Total Damage per Event ($M)"))
title(main = "Weather Events Types in U.S. with\nHighest Total Damage (1950-2011)")
```



Weather Events Types in U.S. with Highest Property Damage (1950−2011)



Weather Events Types in U.S. with Highest Crop Damage (1950−2011)



Weather Events Types in U.S. with Highest Total Damage (1950−2011)