Reproducible Research: Project Assignment 2

Zane Kratzer

Friday, December 19, 2014

The following analysis addresses two specific questions with regards to the U.S. National Oceanic and Atmospheric Administration's (NOAA) Storm Database:

1. Which type of weather events are most harmful with respect to population health?
2. Which type of weather events have the greatest economic consequences?

The following document contains the text, code and output that is used to address these specific questions.

Loading the NOAA Storm Data

The first step is to download the zipped data file and save it to an assigned location.

```
if(!file.exists("./NOAA_Data")){
        dir.create("./NOAA_Data")
}

fileUrl<-"https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
download.file(fileUrl,destfile="./NOAA_Data/StormData.csv.bz2",method="curl")
```

```
## Warning: running command 'curl
```

```
## "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
## -o "./NOAA_Data/StormData.csv.bz2"' had status 127
```

```
## Warning in download.file(fileUrl, destfile =
## "./NOAA_Data/StormData.csv.bz2", : download had nonzero exit status
```

The 'Sys.setlocale' function will help with reading the data into R. The 'read.csv' function is used with the 'bzfile' argument, since the data is in the .bz2 format.

```
Sys.setlocale("LC_ALL","C")
```

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

```
dat<-read.csv(bzfile("./NOAA_Data/StormData.csv.bz2"))
```

Which type of events are most harmful with respect to population health?

I have decided to define "population health" as the combined total number of fatalities and injuries for each event type. The following code adds these field together into a new column called 'pop.health'. The new column is then added to the data frame.

```
pop.health<-dat$FATALITIES + dat$INJURIES
dat<-data.frame(dat,pop.health)
```

The ‘dplyr’ packaged is called and the data is converted into a table for the analysis. The data is grouped by the ‘EVTYPE’ field and the total number of fatalities and injuries for each event type is calculated.

```
## Warning: package 'dplyr' was built under R version 3.1.2
```

```
##
## Attaching package: 'dplyr'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
df<-tbl_df(dat)
sum.dat<-df%>%
        group_by(EVTYPE)%>%
        summarize(sum(pop.health))
```

Additional code is used to clean up the column names and convert the table back into a data frame.

```
colnames(sum.dat)<-c("EventType","PopHealth")
sum.dat<-data.frame(sum.dat)
```

The data set is re-ordered in descending order to help view the event types with the largest number of fatalities and

injuries.

```
sum.dat<-sum.dat[order(sum.dat$PopHealth,decreasing=TRUE),]
sum.dat[1:5,]
```

```
##          EventType PopHealth
## 826        TORNADO     96979
## 124 EXCESSIVE HEAT      8428
## 846      TSTM WIND      7461
## 167          FLOOD      7259
## 453      LIGHTNING      6046
```

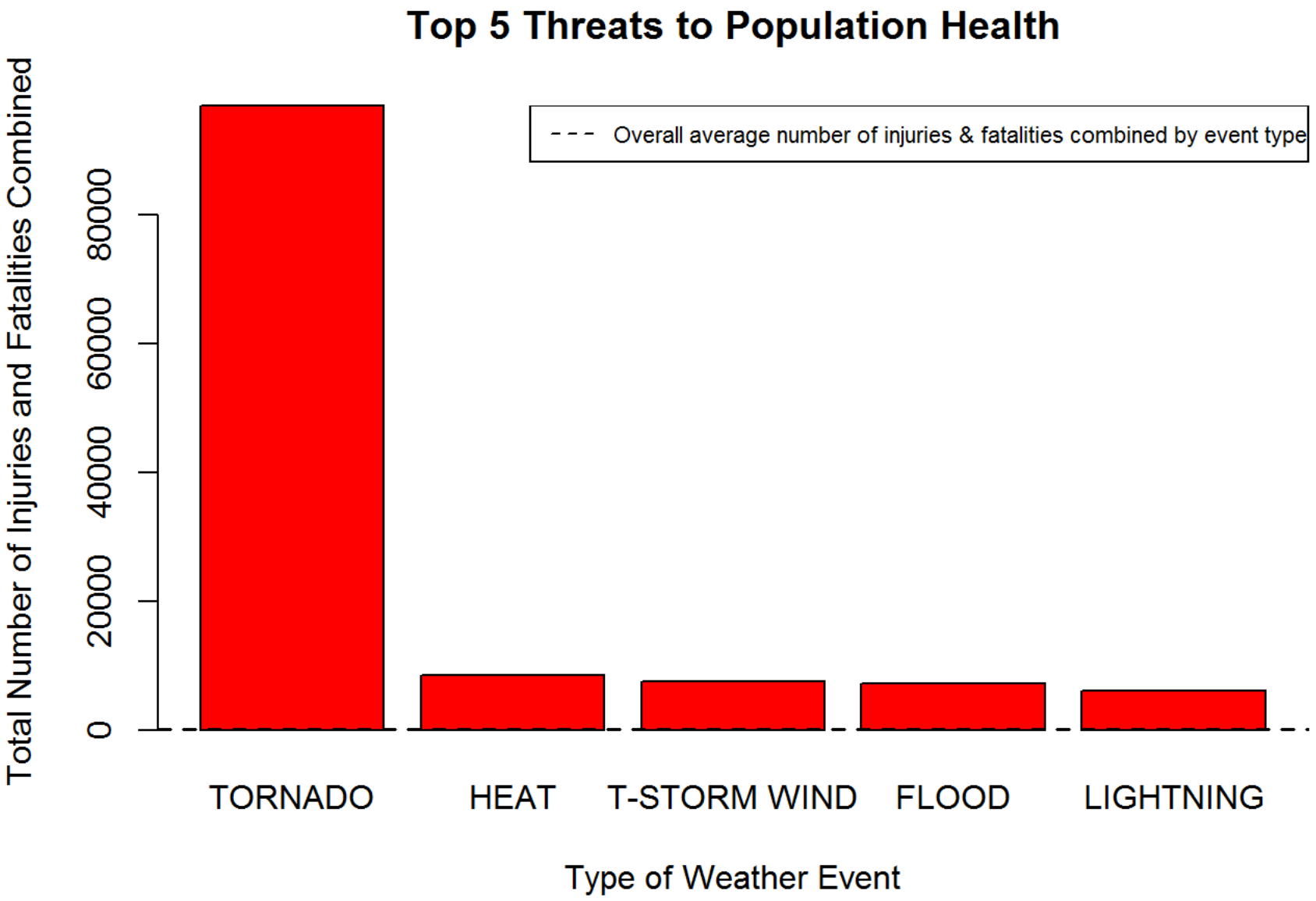
I have decided to pull out the top 5 event types with regards to the total number of fatalities and injuries. Additional code is required to ensure that the event type field is read properly (as a factor variable with only 5 levels now).

```
dat.sub<-sum.dat[1:5,]
dat.sub$EventType<-as.character(dat.sub$EventType)
dat.sub$EventType<-factor(dat.sub$EventType,
                          levels=c("TORNADO","EXCESSIVE HEAT","TSTM WIND","FLOOD","LIGHTNING"))
```

The resulting subset is plotted to show the magnitude of the top 5 weather threats to population health. For comparison purposes, a line is added to the barplot to show the overall average number of injuries and fatalities combined across all event types.

```
barplot(height=dat.sub$PopHealth,
        names.arg=c("TORNADO","HEAT","T-STORM WIND","FLOOD","LIGHTNING"),
        main="Top 5 Threats to Population Health",ylab="Total Number of Injuries and Fatalities Combined",
        xlab="Type of Weather Event",col="red")
```

```
abline(h=mean(sum.dat$PopHealth),lty=2,lwd=1.5)
legend("topright",legend="Overall average number of injuries & fatalities combined by event type",lty=2,lwd
=1,cex=0.7)
```



A review of the data indicates that tornados are by far the largest threat to population health with nearly 100,000 combined fatalities and injuries (N=96,980) during the timeframe of the dataset (1950-2011). The rest of the top 5 weather threats consists of event types that range between 5,000 to 10,000 fatalities and injuries combined: excessive heat, thunderstorm wind, floods and lightning. The average across all event types is very low at only 158. As the plot indicates, this average remains at the very bottom of the plot, suggesting that most weather events should not be considered harmful to the population.

```
summary(sum.dat$PopHealth)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       0       0     158       0   96980
```

From this data, I would recommend that the majority of efforts and resources should be focused on tornado prevention and emergency aid in response to tornados.

Which type of events have the greatest economic consequences?

In order to address the question of economic impact from weather events, the data must be adjusted to reflect the real dollar amounts associated with the costs of property and crop damage. The data consists of 4 fields that will be used for this purpose. Both property and crop damage contains a field with a dollar amount and a field with a multiplier level (K = $1 thoudand, M = $1 million, B = $1 billion).

The following code creates a function that will be used to multiply the dollar amount from the one field with the appropriate level in order to get the real dollar amount for each event.

```
multiplier<-function(x) if(is.null(x) || is.na(x)){1
```

```
        }else if(x == 'K' || x== 'k'){1000
        }else if(x == 'M' || x == 'm'){1000000
        }else if(x == 'B' || x == 'b'){1000000000
        }else 1
```

The function is first vectorized, then it is used to create the 'total_prop_dmg' and 'total_crop_dmg' variables, which reflect the real dollar amounts for each event type. The new 'damage' fields are added to the data frame and it is then saved as a new dataset.

```
vmulti <- Vectorize(multiplier)
total_prop_dmg <- dat$PROPDMG * vmulti(dat$PROPDMGEXP)
total_crop_dmg <- dat$CROPDMG * vmulti(dat$CROPDMGEXP)
dat2<-data.frame(dat,total_prop_dmg,total_crop_dmg)
```

I have decided to define "economic consequences" as the combined costs in property and crop damage for each event type. The following code adds these fields together into a new column called 'total.dmg'. The new column is then added to the data frame.

```
total.dmg<-dat2$total_prop_dmg + dat2$total_crop_dmg
dat2<-data.frame(dat2,total.dmg)
```

The data is converted into a table for the analysis. The data is grouped by the 'EVTYPE' field and the total amount of property and crop damage combined for each event type is calculated.

```
df2<-tbl_df(dat2)
sum.dat2<-df2%>%
        group_by(EVTYPE)%>%
        summarize(sum(total.dmg))
```

Once again, the output data is cleaned up a bit with new column names.

```
colnames(sum.dat2)<-c("EventType","TotalDamage")
sum.dat2<-data.frame(sum.dat2)
```

The data set is re-ordered in descending order to help view the event types with the largest amount of property and crop damage.

```
sum.dat2<-sum.dat2[order(sum.dat2$TotalDamage,decreasing=TRUE),]
sum.dat2[1:5,]
```

```
##             EventType  TotalDamage
## 167             FLOOD 150319678257
## 393 HURRICANE/TYPHOON  71913712800
## 826           TORNADO  57352114049
## 656       STORM SURGE  43323541000
## 241              HAIL  18758221521
```

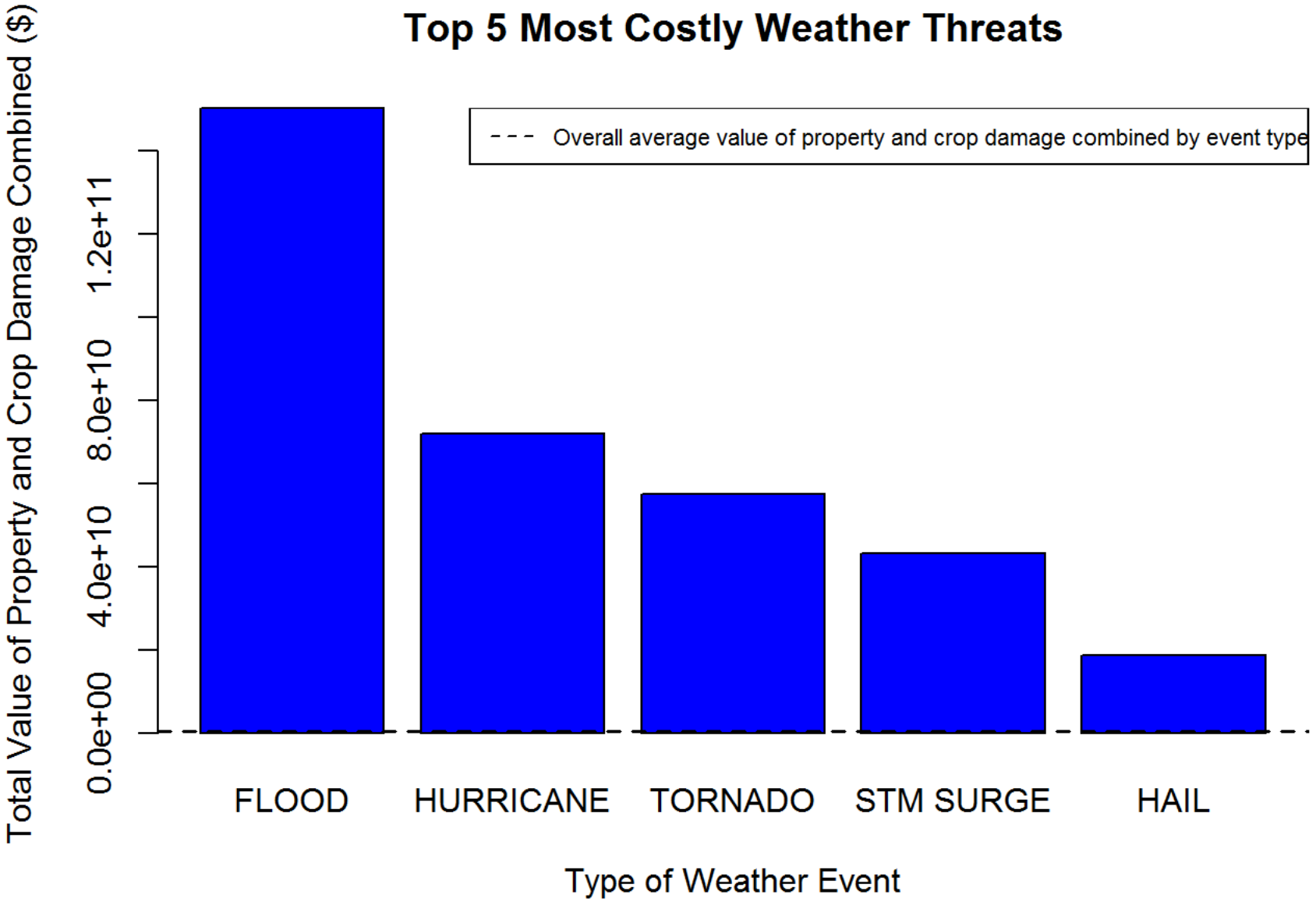
Once again, I have decided to pull out the top 5 event types with regards to the total amount of property and crop damage. Additional code is required to ensure that the event type field is read properly (as a factor variable with only 5 levels now).

```
dat.sub2<-sum.dat2[1:5,]
dat.sub2$EventType<-as.character(dat.sub2$EventType)
dat.sub2$EventType<-factor(dat.sub2$EventType,
                           levels=c("FLOOD","HURRICANE/TYPHOON","TORNADO","STORM SURGE","HAIL"))
```

The resulting subset is plotted to show the magnitude of the top 5 weather threats with regards to economic costs. For

comparison purposes, a line is added to the barplot to show the overall average amount of property and crop damage combined across all event types.

```
barplot(height=dat.sub2$TotalDamage,
        names.arg=c("FLOOD","HURRICANE","TORNADO","STM SURGE","HAIL"),
        main="Top 5 Most Costly Weather Threats",ylab="Total Value of Property and Crop Damage Combined ($)
",
        xlab="Type of Weather Event",col="blue")
abline(h=mean(sum.dat2$TotalDamage),lty=2,lwd=1.5)
legend("topright",legend="Overall average value of property and crop damage combined by event type",lty=2,l
wd=1,cex=0.7)
```



A review of the data indicates that floods have the largest econonmic impact, with over $150 billion dollars worth of damage combined over the full timeframe of the dataset (1950-2011). The rest of the top 5 consists of event types that range between $18 billion to $72 billion worth of total damage: hurricanes/typhoons, tornados, storm surges, and hail. The average across all event types is much lower at only $483.7 million. As the plot indicates, this average remains at the very

bottom of the plot, suggesting that most weather events should not be considered costly in terms of economic consequences.

```
summary(sum.dat2$TotalDamage)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.000e+00 0.000e+00 0.000e+00 4.837e+08 8.500e+04 1.503e+11
```

Although the previous discussion, which considered the impact of weather events on population health, suggested that the majority of resources should be focused on tornado prevention and emergency aid, the second plot indicates that, with concern for economic consequences, the top 4 weather events at the least (floods, hurricanes/typhoons, tornados, and storm surges) all deserve a significant amount of attention and resouces based on the total costs in property and crop damage that these weather events have inflicted over the course of the dataset (ranging from $43 billion for storm surges to $150 billion for floods).