Project2.Rmd

June 2015

Analysis of the U.S. National Weather Service Storm Data from 1950-2011: Health and Economic Impacts

Synopsis:

Data from the National Weather Service Storm Data from 1950 to November 2011 was analyzed for the 50 U.S. states to determine which storm events produced the most economic impact and health risks.

Data Processing:

Data was downloaded from the Coursera Reproducible Research project page: repdata_data_StormData.csv.bz2 The 'gunzip repdata_data_StormData.csv.bz2 command was used to uncompress the file at the command line, rather than unzipping within R (to reduce analysis time). To include the uncompression of the "original" source file, uncomment the command, *data.file <- bzfile("repdata_data_StormData.csv.bz")*. The resulting repdata_data_StormData.csv file was then subset to include only the 50 U.S. states (ignoring the U.S. territories) and ignoring any NA's.

Two separate data frames were created, one for the health impacts and another for economic impacts. The health impact analysis involved plotting the fatalities and injuries as a function of the event type. The economic impact analysis involved generating a plot of the average cost (in U.S. dollars) in property damage and in crop damage as a function of event type. The cost in U.S. dollars was calculated by converting the 'K', 'M', and 'B' in the PROPDMGEXP or CROPDMGEXP column into the corresponding multiplier; 1000 for K, 1000000 for M, 100000000 for B and 1 for anything else. These multipliers were then added to the corresponding PROPDMG or CROPDMG value. A plot of the cost in U.S. dollars vs the event type was then generated to show which events have the most economic impact. The data was averaged over the 1950-November 2011 period (the currently available data).

Too many events made the resulting bar plots too difficult to read, therefore further subsetting of the health impact and economic impact data frames was needed. Initially, the mean number of health impacts (sum of fatalities and injuries) and the mean number of damage (sum of property damage and crop damage) were used as criteria for subsetting the health data. The resulting plots were still too cluttered. Next, the top 20th value for the total cost (sum of property and crop damage) and total number of health impacts (sum of fatalities and injuries) were used to subset the data.

Software

- MacOS 10.10.3
- 2.53 GHz Intel Core 2 Duo
- Memory: 4 GB 1067 MHz DDRS

- RStudio version .99.442
- R version 3.2.0 “Full of Ingredients”

Packages

- dplyr
- reshape2
- ggplot2

Code:

```
      library(dplyr)
```

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

```
      library(reshape2)
      library(ggplot2)

      #Set the working diretory
      setwd("~/Coursera/ReproducibleResearch/Project2")

      #unzip and open the csv file
      #For performance, uncompress from the command line and use the .csv file, ot
herwise,
      #uncomment the line below, and comment the the data.file<-"repdata_data_Stor
mData.csv"
      #if you wish to start from the original, compressed file.
      #data.file <- bzfile("repdata_data_StormData.csv.bz")
      data.file <- "repdata_data_StormData.csv"
      raw.data <- read.csv(data.file, sep=",", header=TRUE, stringsAsFactors=FALSE
)

      #Replace any NAs with 0
      raw.data[is.na(raw.data)] <- 0


      #Subset data to the columns that are pertinent to our analysis:
      #event type (EVTYPE), property damage (PROPDMG, PROPDMGEXP), crop
      #damage(CROPDMG, CROPDMGEXP), fatalities (FATALITIES), injuries (INJURIES).
      #We will also focus on just the 50 states (omitting the U.S. territories)
      #and we will use the begin date to get the year information, so we can
      #look at the trends through the years.

      # use dplyr to select the relevant columns
      selected.data <- select(raw.data,BGN_DATE,STATE,EVTYPE,PROPDMG,PROPDMGEXP,CR
OPDMG,CROPDMGEXP,
                              FATALITIES,INJURIES)

      #Subset for the 50 states. Get a list of the unique states, the first 50
      #are the ones we want.
      states <- unique(selected.data$STATE)
      us <- head(states,50)
      us.only <- filter(selected.data, STATE %in% us)

      #Convert the BGN_DATE string to a Date type and obtain the year.
      # We may want to use the year if we plan on doing an analysis that
      # looks at the data over each year.
      datetimes <- strptime(us.only$BGN_DATE, "%m/%d/%Y %H:%M:%S")
      years <- format(datetimes,format="%Y")
      years <- as.numeric(years)
      us.only <- mutate(us.only, YEAR=c(years))

      #Convert the FATALITIES and INJURIES to numeric types so we can
      #calculate sums and means.
```

```
      fatalities <- as.numeric(us.only$FATALITIES)
      injuries <- as.numeric(us.only$INJURIES)
      us.only$FATALITIES <- fatalities
      us.only$INJURIES <- injuries

      #Convert the PROPDMG and CROPDMG columns to numeric types so we can
      #calculate the dollar value of damage.
      props <- as.numeric(us.only$PROPDMG)
      crops <- as.numeric(us.only$CROPDMG)
      us.only$PROPDMG <- props
      us.only$CROPDMG <- crops

      #Convert the codes in the PROPDMGEXP and CROPDMGEXP columns into numeric
      #values.  'K' = 1000, 'M' = 1,000,000 , 'B' = 1,000,000,000 and blank
      #is set to 1.  Create a vector of these values, append them to the
      #us.only dataframe and then multiply the PROPDMG/CROPDMG to the correspondin
g
      #PROPDMGEXP/CROPDMGEXP column to obtain the total cost in U.S. dollars.
      #Save these results to a new column in us.only, named PROPCOST and CROPCOST,
      #respectively.
      prop.multiplier <- c()
      for(i in 1:length(us.only$PROPDMG)){
            if( us.only[i,5] == 'B'){
                  #Set billions
                  prop.multiplier <- append(prop.multiplier, c(1000000000))
            }
            else if( us.only[i,5] == 'M'){
                  #Set millions
                  prop.multiplier <- append(prop.multiplier, c(1000000))
            }
            else if( us.only[i,5] == 'K'){
                  #Set thousands
                  prop.multiplier <- append(prop.multiplier, c(1000))
            }
            else{
                  #Blank, this is set to 1, so we will use the actual
                  #amount in the PROPDMG column as is.
                  prop.multiplier <- append(prop.multiplier, c(1))
            }

      }

      #Repeat for the CROPDMGEXP column
      crop.multiplier <- c()

      for(i in 1:length(us.only$CROPDMG)){
            if( us.only[i,7] == 'B'){
                  #Set billions
                  crop.multiplier <- append(crop.multiplier, c(1000000000))
```

```
            }
            else if( us.only[i,7] == 'M'){
                  #Set millions
                  crop.multiplier <- append(crop.multiplier, c(1000000))
            }
            else if( us.only[i,7] == 'K'){
                  #Set thousands
                  crop.multiplier <- append(crop.multiplier, c(1000))
            }
            else{
                  #Blank, this is set to 1, so we will use the actual
                  #amount in the PROPDMG column as is.
                  crop.multiplier <- append(crop.multiplier, c(1))
            }

      }

      #Make the crop.multiplier a numeric vector before we add it to the us.only
      #data frame
      crop.mult <- as.numeric(crop.multiplier)

      #Make the prop.multiplier a numeric vector before we add it to the us.only
      #data frame
      prop.mult <- as.numeric(prop.multiplier)

      # use dplyr to create new columns for the calculated cost of damage.
      us.only <- mutate(us.only, PROPMULT=c(prop.mult))
      us.only <- mutate(us.only, PROPCOST=PROPDMG*PROPMULT)
      us.only <- mutate(us.only, CROPMULT=c(crop.mult))
      us.only <- mutate(us.only, CROPCOST=CROPDMG*CROPMULT)

      #Calculte the totals for property damage and health impacts. These
      #will be used to subset the data further.
      us.only <- mutate(us.only, TOTALCOST=CROPCOST+PROPCOST)
      us.only <- mutate(us.only, TOTALHEALTH=FATALITIES+INJURIES)

      #Create new dataframes for damages and health based on the relevant
      #data.
      us.data <- select(us.only, YEAR, EVTYPE, PROPCOST, CROPCOST, FATALITIES,INJU
RIES )
      us.damages <- select(us.only, EVTYPE, PROPCOST,CROPCOST,TOTALCOST)
      us.health <- select(us.only,  EVTYPE, FATALITIES, INJURIES,TOTALHEALTH)

      #Now generate plots to answer question1: Which event type(s) are the most
      #harmful to U.S. human health?
      #In other words, over the entire data set (all years available and the 50
      #U.S. states only), what are the average/mean number
      #of fatalities and what are the mean number of injuries for each
      #event type?
```

```
      # First, melt the data, then cast so that we can readily calculate the
      # average number of Fatalities and the average number of Injuries for
      # each event type. Subset data so we don't get overwhelmed with too
      # many event types, therefore, only include the top 20 events.
      sorted.health <- sort(us.health$TOTALHEALTH,decreasing=TRUE)
      top20.health <- sorted.health[20]
      us.health.sub <- subset(us.health, TOTALHEALTH >= top20.health)

      melted.us.health <- melt(us.health.sub, id=c("EVTYPE"),measure=c("FATALITIES
","INJURIES"))
      mean.us.health <- dcast(melted.us.health,EVTYPE~variable,mean)
      mean.us.health.long <- melt(mean.us.health)
```

```
## Using EVTYPE as id variables
```

```
      #Generate the bar plot of the fatalities and injuries based on event.
      png(file="./health_impact.png",width=480, height=480)
      p1<-ggplot(mean.us.health.long, aes(EVTYPE,value,fill=variable))+
            geom_bar(stat="identity",position="dodge")+
            theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))+
            xlab("Event Type") + ylab("Number of people") +
            ggtitle("Health impacts of U.S. Weather Events averaged from 1950-2011
")+
            coord_flip()
      print(p1)
      dev.off()
```

```
## quartz_off_screen
##                 2
```

```
    #Answer question 2, which event types have the most economic impact?
    #Generate a plot of the event type vs. cost, broken down by property
    #damage and crop damage.
    # First, melt the data, then cast so that we can readily calculate the
    # average number of Fatalities and the average number of Injuries for
    # each event type.

    #Sort the damage costs and use the top 20
    #as a criteria for subsetting the data.
    sorted.damages <- sort(us.damages$TOTALCOST,decreasing=TRUE)
    top20.damages <- sorted.damages[20]
    us.damages.sub <- subset(us.damages, TOTALCOST >= top20.damages)

    melted.us.damage <- melt(us.damages.sub, id=c("EVTYPE"),measure=c("PROPCOST","
CROPCOST"))
    mean.us.damage <- dcast(melted.us.damage,EVTYPE~variable,mean)
    mean.us.damage.long <- melt(mean.us.damage)
```

```
## Using EVTYPE as id variables
```

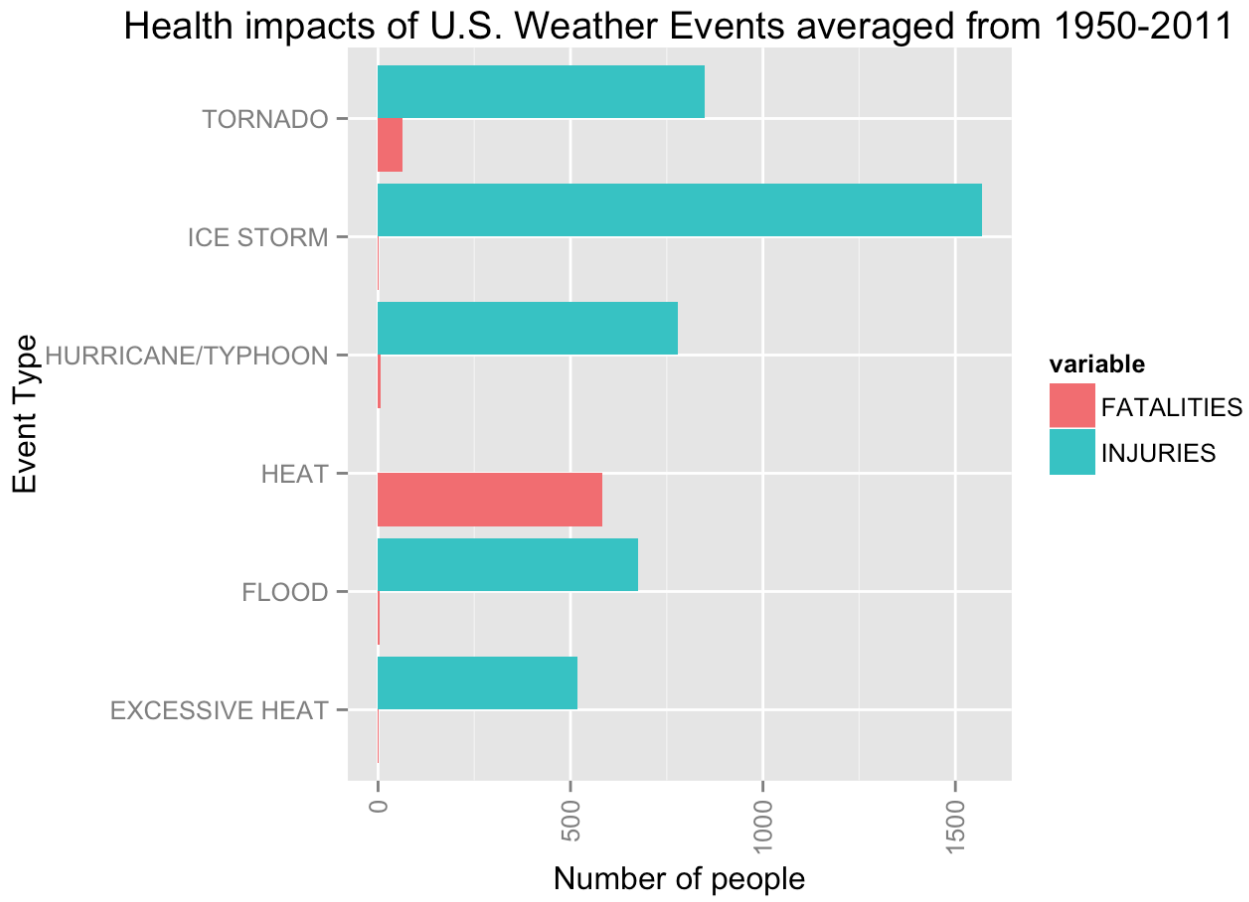
```
    #Format the property cost and crop cost values so ggplot2 doesn't
    #generate warnings about changing widths for bar size.
    formatted.propcost <- format(mean.us.damage.long$PROPCOST, digits=2)
    formatted.cropcost <- format(mean.us.damage.long$CROPCOST, digits=2)

    #Generate the bar plot of property cost and crop cost due to event.
    png(file="./economic_impact.png",width=480,height=480)
    p2 <-ggplot(mean.us.damage.long, aes(EVTYPE,value,fill=variable))+
          geom_bar(stat="identity",position="dodge")+
          theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))+
          xlab("Event Type") + ylab("Cost in U.S. Dollars") +
          ggtitle("Economic Impact of U.S. Weather Events averaged from 1950-2011"
)+
          coord_flip()
    print(p2)
    dev.off()
```

```
## quartz_off_screen
##                 2
```

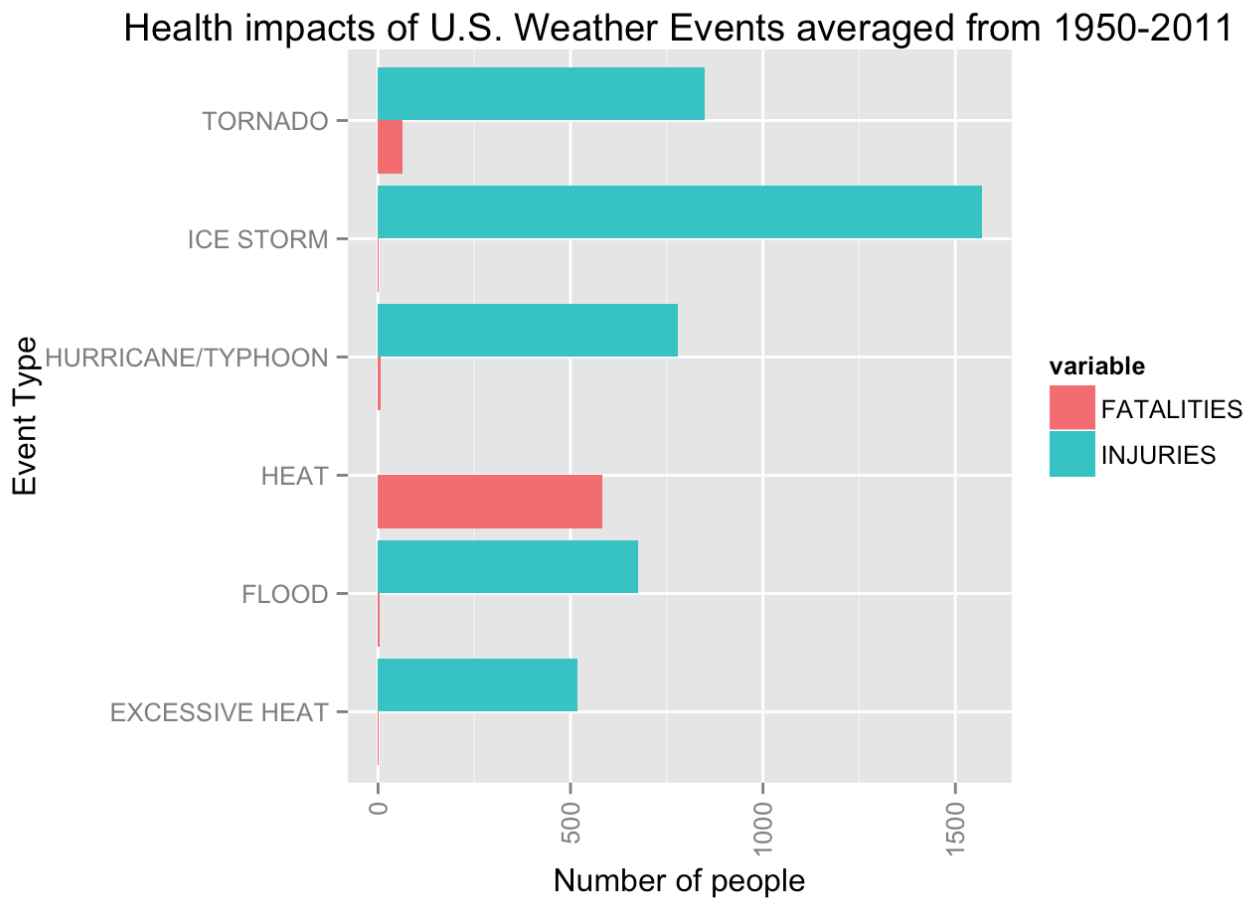
```
    #Summary of the max average property damage, crop damage, number of fatalities
, and number of
    #injuries
    max.property <- max(mean.us.damage$PROPCOST)
    max.crop <- max(mean.us.damage$CROPCOST)
    max.fatalities <- max(mean.us.health$FATALITIES)
    max.injuries <- max(mean.us.health$INJURIES)
```

Results

```
    # Generate bar plot for Economic impacts.
    p1<-ggplot(mean.us.health.long, aes(EVTYPE,value,fill=variable))+
            geom_bar(stat="identity",position="dodge")+
            theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))+
            xlab("Event Type") + ylab("Number of people") +
            ggtitle("Health impacts of U.S. Weather Events averaged from 1950-2011
")+
            coord_flip()
    print(p1)
```



```
    # Generate bar plot for Health impacts.
    p1<-ggplot(mean.us.health.long, aes(EVTYPE,value,fill=variable))+
            geom_bar(stat="identity",position="dodge")+
            theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))+
            xlab("Event Type") + ylab("Number of people") +
            ggtitle("Health impacts of U.S. Weather Events averaged from 1950-2011
")+
            coord_flip()
    print(p1)
```



```
    #Summary of maximum property damage, crop damage, number of fatalities, and nu
mber of injuries
    #for the 50 U.S. states averaged over the 1950-November 2011 data set.
    max.property.event.index <- which.max(mean.us.damage$PROPCOST)
    max.property.event <- mean.us.damage[max.property.event.index,1]
    max.property.cost <- max(mean.us.damage$PROPCOST)

    max.crop.event.index <- which.max(mean.us.damage$CROPCOST)
    max.crop.event <- mean.us.damage[max.crop.event.index,1]
    max.crop.cost <- max(mean.us.damage$CROPCOST)

    max.fatalities.index <- which.max(mean.us.health$FATALITIES)
    max.fatalities.event <- mean.us.health[max.fatalities.index,1]
    max.fatalities.number <- max(mean.us.health$FATALITIES)

    max.injuries.index <-which.max(mean.us.health$INJURIES)
    max.injuries.event <- mean.us.health[max.injuries.index,1]
    max.injuries.number <- max(mean.us.health$INJURIES)
```

Summary of maximum property damage, crop damage, number of fatalities, number of injuries (*averaged over the 50 U.S. states, for 1950-2011*):

Maximum property damage (in U.S. Dollars): 5.910^{10} **caused by** FLOOD
Maximum crop damage (in U.S. Dollars): 510^{9} **caused by** ICE STORM
Maximum number of fatalities: 583 **caused by** HEAT
Maximum number of injuries: 1568 **caused by** ICE STORM

The plot of economic impacts shows that floods, storm surge, hurricane/typhoon, winter storms, and tropical storms accounted for the majority of property damage. Floods caused the most property damage, followed by storm surge. River flooding and ice storms were the two events resulting in the most crop damage.

The plot of the health impacts indicate that tropical storms, excessive heat, tornadoes, and flash floods were responsible for the most fatalities. Excessive heat, winter weather mix, and tornadoes were the top three events resulting in the most number of injuries.

Further Work

A further analysis of storm events broken down by year or decades might provide agency managers with annual or decadal trends which could be used for longer term planning and budgeting. A more regional study where the cost and type of storm events are broken down by region might prove useful for local/state governments in planning efforts.