Impact of weather events on the population health and property/crop damage

Rok Bohinc

June 15, 2019

This project involves exploring the U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database. This database tracks characteristics of major storms and weather events in the United States, including when and where they occur, as well as estimates of any fatalities, injuries, and property damage.

In this study I adress the following questions:

1. Across the United States, which types of events are most harmful with respect to population health?
2. Across the United States, which types of events have the greatest economic consequences?

In order to address this questions I first subset the data to look at only the relevant variables and filter the data for the recent present events. Then I make the data set tidy and group the event types for easier data analysis. I then investigate which weather events cause the most injuries and fatalities/crop and property damage. I further investigate injuries and fatalities of tornadoes by looking at the dependence on the intensity scale. I also look where and when the most damaging weather events occour and find possible explanations for the observations.

Data Procesing

First of all I load the data for the project.

```
setwd("/home/rok/Edjucation/2019.3.28. Data_Science-Specialization/Reproducible research/Final Assighnme
if (!file.exists("data")){dir.create("data")}
fileURL <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"

# download data

download.file(fileURL, destfile = "./data/Storm.csv", method = "curl")

# record the date the data has been downloaded

dateDownloaded <- date()
data <- read.csv("./data/Storm.csv")
```

The data has 902297 observations and 37 variables, which is a lot. On top of this the data is quite messy, so I will try to reduce the number of rows and columns of the data set.

Subsetting data

In order to answer both of the questions not all of the 37 variables are needed. It therefore makes sense to subset the data by choosing only the relevant variables including the state where and when the event has occurred, the type of event, the Fujita tornado intensity scale, the magnitude of the event, the associated fatalities, injuries, property damage and crop damage.

```
library(dplyr)
data <- select(data, STATE__, BGN_DATE, EVTYPE, F, MAG, FATALITIES, INJURIES, PROPDMG, PROPDMGEXP, CROPD
```

Both of the questions are associated with the **current** impact of storm and weather events on the population. So my aim is to filter the data for the "recent" present. First I convert the data column to the date format.

```
library(lubridate)
data[,2] <- mdy_hms(data[,2])
period <- 10
```

I see that the data describe events from 1950-01-03 to 2011-11-30. I make the assumption that events that happened in the last 10 years since 2011-11-30 represent the "recent"" present.

```
data <- filter(data, BGN_DATE >= last(data$BGN_DATE) - years(period))
```

The reduced data set has 454775 observations and 11 variables.

Furthermore, to answer both of the question I can restrict myself to only look at fatalities, injuries, crop damage and property damage different from 0.

```
data <- subset(data, (FATALITIES != 0 | INJURIES != 0) | (PROPDMG != 0 | CROPDMG !=0))
```

With this I have reduced the number if observations to 134763.

Cleaning data

First of all I will convert property damage and crop damage both to a single column. In the original data this information is composed of the XXXDMG number and the XXXDMGEXP, where XXX corresponds either to CROP or PROP. The exponent contains the following values:

```
table(as.factor(as.character(data$PROPDMGEXP)))
```

```
##
##             B      K      M
##   3203     27 126609   4924
```

```
table(as.factor(as.character(data$CROPDMGEXP)))
```

```
##
##           B     K     M
## 46693     2 87050  1018
```

where K stands for thousand, M for million, and B for billion. I now first create two new column for the exponents where I make the transformation and then afterwards I create another two columns multiplying XXXDMG with the appropriate factor.

```
data <- mutate(data, propdamageexp = 0, cropdamageexp = 0)

data$propdamageexp[data$PROPDMGEXP=="K"] <- 1000
data$propdamageexp[data$PROPDMGEXP=="M"] <- 1000000
data$propdamageexp[data$PROPDMGEXP=="B"] <- 1000000000

data$cropdamageexp[data$CROPDMGEXP=="K"] <- 1000
data$cropdamageexp[data$CROPDMGEXP=="M"] <- 1000000
data$cropdamageexp[data$CROPDMGEXP=="B"] <- 1000000000

data <- mutate(data, propdamage = PROPDMG*propdamageexp, cropdamage = CROPDMG*cropdamageexp)
data <- select(data, STATE__, BGN_DATE, EVTYPE, F, MAG, FATALITIES, INJURIES, propdamage, cropdamage)
```

Making the data set tidy

Now I want to create a tidy data set. In particular I want to gather together information about injuries and fatalities, and the information about the property damage and crop damage.

```
library(tidyr)
cleandata <- gather(data, damagetype, value, c(propdamage, cropdamage)) %>% gather(injurytype, count, c
names(cleandata) <- c("statefips", "date", "eventtype", "tornadointscale", "magnitude", "damagetype", "
cleandata <- droplevels(cleandata) # Drop unused factors
```

Grouping of event types

The last part in the data processing is grouping the event types. As you can see below there are several event types that can be grouped together, as for instance STRONG WIND and STRONG WINDS.

```
head(sort(unique(cleandata$eventtype)))
```

```
## [1]    HIGH SURF ADVISORY  ASTRONOMICAL HIGH TIDE ASTRONOMICAL LOW TIDE
## [4] AVALANCHE              BLIZZARD               COASTAL FLOOD
## 81 Levels:    HIGH SURF ADVISORY ... WINTER WEATHER/MIX
```

Below is the conversion/simplification of event types. This is a rather unrigorous conversion and certain classifications can certainty be classified otherwise.

```
cleandata$eventtype <- as.character(cleandata$eventtype)
cleandata$eventtype[grep("STORM|HURRICANE|BLIZZARD|FUNNEL CLOUD",cleandata$eventtype)] <- "STORM"
cleandata$eventtype[grep("WIND|DUST DEVIL|DRY MICROBURST|TROPICAL DEPRESSION",cleandata$eventtype)] <-
cleandata$eventtype[grep("FLOOD|FLD",cleandata$eventtype)] <- "FLOOD"
cleandata$eventtype[grep("FIRE|DENSE SMOKE",cleandata$eventtype)] <- "FIRE"
cleandata$eventtype[grep("SNOW",cleandata$eventtype)] <- "SNOW"
cleandata$eventtype[grep("SURF",cleandata$eventtype)] <- "SURF"
cleandata$eventtype[grep("RAIN|PRECIPITATION",cleandata$eventtype)] <- "RAIN"
cleandata$eventtype[grep("TIDE",cleandata$eventtype)] <- "TIDE"
cleandata$eventtype[grep("COLD",cleandata$eventtype)] <- "COLD"
cleandata$eventtype[grep("HEAT",cleandata$eventtype)] <- "HEAT"
cleandata$eventtype[grep("CURRENT",cleandata$eventtype)] <- "CURRENT"
cleandata$eventtype[grep("WINTER",cleandata$eventtype)] <- "WINTER"
cleandata$eventtype[grep("FOG",cleandata$eventtype)] <- "FOG"
cleandata$eventtype[grep("FROST|ICE ON ROAD",cleandata$eventtype)] <- "FROST"
cleandata$eventtype[grep("HAIL",cleandata$eventtype)] <- "HAIL"
cleandata$eventtype[grep("TORNADO|WATERSPOUT",cleandata$eventtype)] <- "TORNADO"
cleandata$eventtype[grep("SEA",cleandata$eventtype)] <- "SEA"

# Making factor variables
cleandata$eventtype <- as.factor(cleandata$eventtype)
cleandata$tornadointscale <- as.factor(cleandata$tornadointscale)
cleandata$injurytype <- as.factor(cleandata$injurytype)
```

I have reduced the eventtype variable to 24 different entries. In the end I show the head of the processed data set I named cleandata.

```
head(cleandata)
```

```
##   statefips       date eventtype tornadointscale magnitude damagetype
## 1         1 2001-11-29      WIND            <NA>        50 propdamage
## 2         1 2001-11-29      WIND            <NA>        50 propdamage
```

```
## 3         1 2001-12-14      WIND            <NA>        60 propdamage
## 4         1 2001-12-14   TORNADO               0         0 propdamage
## 5         1 2001-12-17      WIND            <NA>        50 propdamage
## 6         2 2001-12-07 AVALANCHE            <NA>         0 propdamage
##   value injurytype count
## 1 25000 FATALITIES     0
## 2  2000 FATALITIES     0
## 3 25000 FATALITIES     0
## 4 12000 FATALITIES     0
## 5  8000 FATALITIES     0
## 6  4000 FATALITIES     0
```

Results

General overview

Below is the code for investigating the dependence of fatalities, injuries, crop damage and property damage on the event type.

```
summeddataFI <- group_by(cleandata, eventtype, injurytype) %>% summarise(sum(count)) %>% arrange(desc(`s
summeddataDMG <- group_by(cleandata, eventtype, damagetype) %>% summarise(sum(value)) %>% arrange(desc(`

par(mfrow=c(4,1))
barplot(subset(summeddataFI, injurytype=="FATALITIES")$`sum(count)`, names.arg = subset(summeddataFI, in
barplot(subset(summeddataFI, injurytype=="INJURIES")$`sum(count)`, names.arg = subset(summeddataFI, inju
barplot(subset(summeddataDMG, damagetype=="cropdamage")$`sum(value)`, names.arg = subset(summeddataDMG,
barplot(subset(summeddataDMG, damagetype=="propdamage")$`sum(value)`, names.arg = subset(summeddataDMG,
```

The most injuries and fatalities have been caused by tornadoes. The second most injuries and fatalities have been caused by events related to heat. Not surprisingly the most crop damage has been caused by drought, flood, and storms. In 10 years since 2011-11-30 drought, flood, and storms have caused about 11.7, 8.8, and 7.8 billion USD of crop damage, respectively. On the other hand drought does not damage the property as severely and most damage to properties is done though flood and storms. In 10 years since 2011-11-30 flood and storms have caused about 288.7, and 258 billion USD of property damage.

Injuries and fatalaties caused by tornados

Below I show a histogram plot of tornado fatalities and injuries sorted according to the Fujita tornado intensity scale.

```
library(ggplot2)
library(gridExtra)

tornadodata <-filter(subset(cleandata, eventtype=="TORNADO"),!is.na(tornadointscale),count!=0)

p1  <- qplot(count, data = filter(tornadodata, injurytype=="FATALITIES"), fill = tornadointscale, bins=5

p2  <- qplot(count, data = filter(tornadodata, injurytype=="INJURIES"), fill = tornadointscale, bins=10

grid.arrange(p1, p2, nrow = 1)
```

We can see that as the tornado intensity scale number (0-5) increases so does the number of fatalities/injuries. Above I have restricted the plot to a specific range leaving out outlines. However the maximum amount of

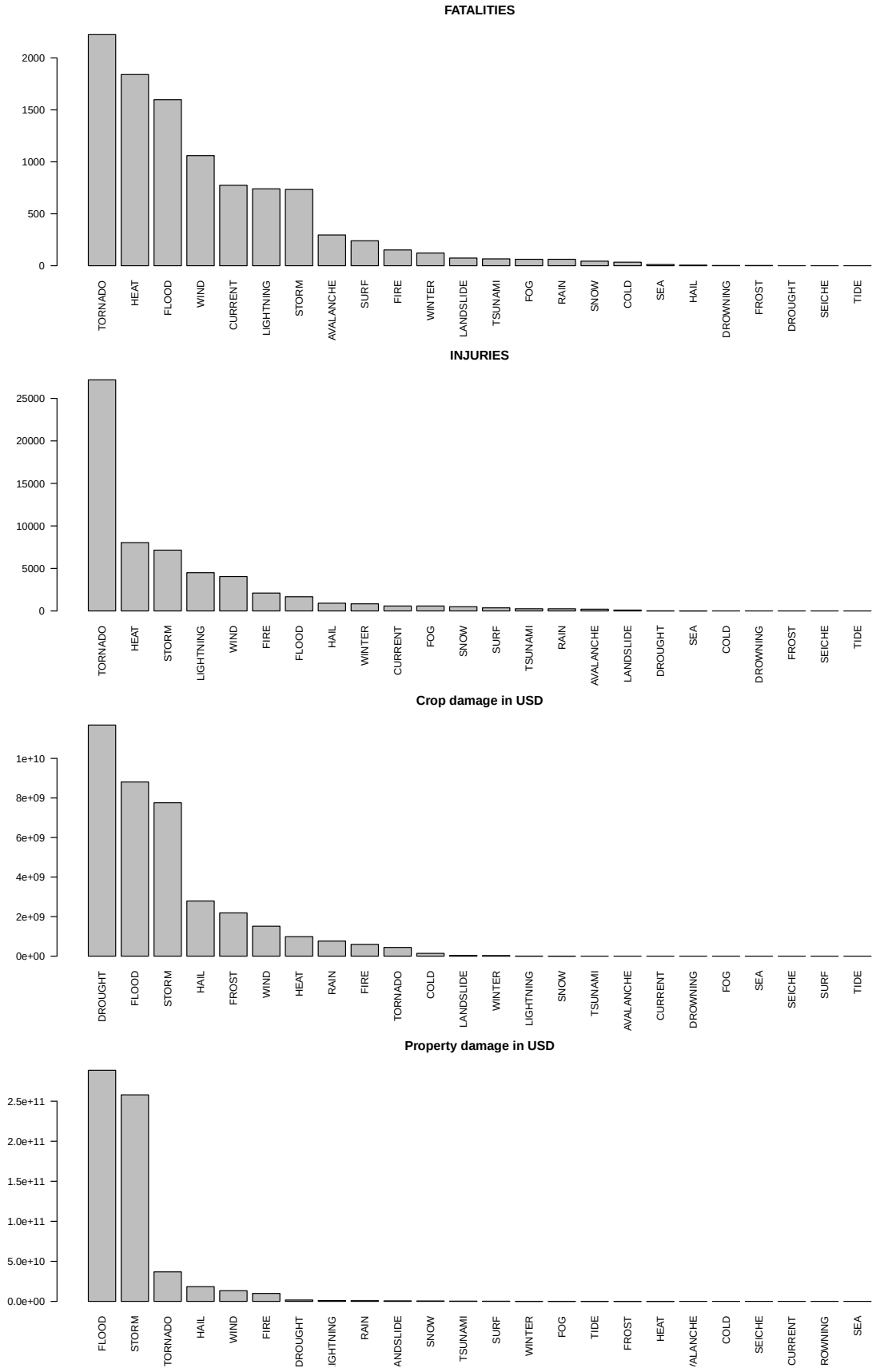

Figure 1: Figure: This plot shows the number of fatalaties and injuries, as well as crop damage and property damage across different event types.

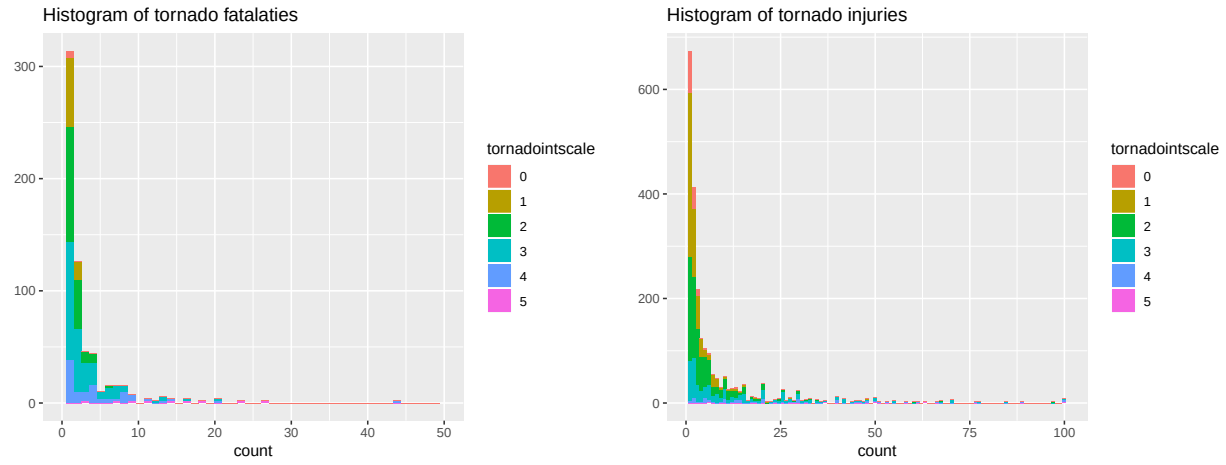

Figure 2: Figure: Histogram of tornado fatalities and injuries sorted according to the Fujita tornado intensity scale.

injuries/fatalities in an tornado event are 1150 and 158, respectively, both occurring in a tornado level 5 in the intensity scale.

```
group_by(tornadodata,tornadointscale, injurytype) %>% summarise(sum(count)) %>% arrange(injurytype, des
```

```
## # A tibble: 12 x 3
## # Groups:   tornadointscale [6]
##    tornadointscale injurytype `sum(count)`
##    <fct>           <fct>             <dbl>
##  1 3               FATALITIES          746
##  2 5               FATALITIES          572
##  3 4               FATALITIES          542
##  4 2               FATALITIES          264
##  5 1               FATALITIES           94
##  6 0               FATALITIES            6
##  7 3               INJURIES           9618
##  8 4               INJURIES           7922
##  9 2               INJURIES           4546
## 10 5               INJURIES           3036
## 11 1               INJURIES           1810
## 12 0               INJURIES            248
```

```
sort(table(tornadodata$tornadointscale), decreasing = TRUE)
```

```
##
##   2   3   1   4   0   5
## 976 784 712 244 146  34
```

We can see that most fatalities (746) and injuries (9618) have occurred in tornadoes of level 3. This is partially related to the fact that such tornadoes are the second most frequent (784 events). Interestingly there were 572 fatalities in level 5 tornadoes (second most) although the were only 34 such tornadoes recorded in the 10 years time period.

Tornado fatalities and injuries by states

```
library(tidyr)
tornadostate <- group_by(tornadodata, statefips, injurytype) %>% summarise(sum(count)) %>% arrange(inju
tornadostate <- as.data.frame(unclass(tornadostate))
spread(tornadostate, injurytype, sum.count.) %>% arrange(desc(FATALITIES), desc(INJURIES)) %>% head
```

```
##   statefips FATALITIES INJURIES
## 1         1        548     5070
## 2        29        464     4004
## 3        47        320     3244
## 4        28         98     1150
## 5         5         92     1176
## 6        13         86     1174
```

From the table above we can see that states most prone to tornado fatalities and injuries are Alabama (01), MISSOURI (29), and TENNESSEE (47), which are all neighboring states in the central/southeast part of the US.

Total damage by states

Here I am going to focus on which states are mostly affected by drought, flood, and storms.

```
stormdata <- filter(cleandata, eventtype=="STORM")
flooddata <- filter(cleandata, eventtype=="FLOOD")
droughtdata <- filter(cleandata,eventtype=="DROUGHT")

group_by(flooddata, statefips) %>% summarise(sum(value)) %>% arrange(desc(`sum(value)`)) %>% head
```

```
## # A tibble: 6 x 2
##   statefips `sum(value)`
##       <dbl>        <dbl>
## 1         6 232414482000
## 2        47   9108424340
## 3        19   5734830000
## 4        36   5276415780
## 5        34   4015980000
## 6        42   3594593000
```

```
group_by(stormdata, statefips) %>% summarise(sum(value)) %>% arrange(desc(`sum(value)`)) %>% head
```

```
## # A tibble: 6 x 2
##   statefips `sum(value)`
##       <dbl>        <dbl>
## 1        22 110903790200
## 2        12  58582372400
## 3        28  52841726320
## 4        48  17810012100
## 5         1   7476859800
## 6        40   2842034500
```

```
group_by(droughtdata, statefips) %>% summarise(sum(value)) %>% arrange(desc(`sum(value)`)) %>% head
```

```
## # A tibble: 6 x 2
##   statefips `sum(value)`
##       <dbl>        <dbl>
```

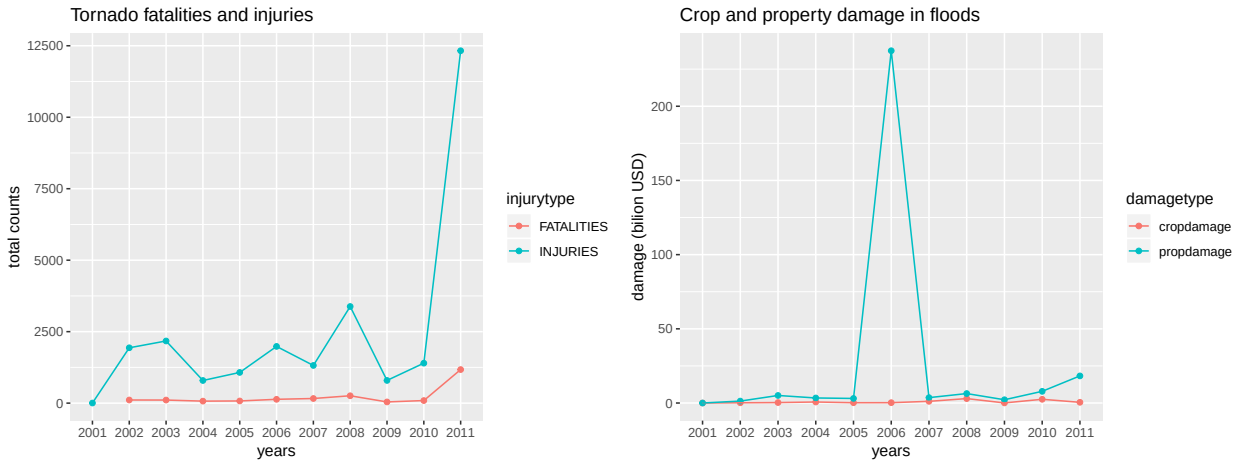

Figure 3: Figure: Time series plot of total fatalities/injuries (left) and property/crop damage (right) from 2001 to 2011.

```
## 1         48   7750774000
## 2         19   2309940000
## 3         31    960000000
## 4         13    688000000
## 5         17    569140000
## 6         40    459738000
```

From here we can see that the flood has done by far most damage in California (06), storms have done most damage in Louisiana (22), followed by Florida (12) and Mississippi (28), while drought has caused most damage in Texas (48), which is one of the hottest states in the US.

Time analysis

Here I am going to focus on tornadoes and flood has influenced injuries/fatalities and damage during the last 10 years.

```
library(ggplot2)
library(gridExtra)
tornadodata$date <- year(tornadodata$date)
timetornado <- group_by(tornadodata, date, injurytype) %>% summarise(sum(count))

flooddata$date <- year(flooddata$date)
timeflood <- group_by(flooddata, date, damagetype) %>% summarise(sum(value))

p1 <- qplot(data = timetornado, date, `sum(count)`,  geom=c("point", "line"), ylab = "total counts", ma

p2 <- qplot(data = timeflood, date, `sum(value)`/1000000000, ylab = "damage (bilion USD)", main = "Crop
grid.arrange(p1, p2, nrow = 1)
```

So we can see that the number of injuries and fatalities is increasing over the years. In particularly in 2011 there was substantially more injuries and deaths compared to previous years. The observed increase could be caused by global warming. On the right side plot we see that property damage was substantially larger in 2006 compared to other years.

```
group_by(filter(flooddata, date==2006, damagetype=="propdamage"), statefips) %>% summarise(sum(value)) %
```

```
## # A tibble: 53 x 2
##    statefips `sum(value)`
##        <dbl>        <dbl>
##  1         6 230606182000
##  2        36   1977120000
##  3        22   1299654000
##  4        39   1099517600
##  5        42   1063060000
##  6        48    438524000
##  7        51    144645000
##  8         2    122444680
##  9        41     98000000
## 10        34     67000000
## # ... with 43 more rows
```

From the table above we can see that most of the damage has occurred in California (06). By looking into the web I have found that there was a flood in California 2005/2006, which likely caused the huge property damage.