Outcomes of severe weather events in the USA a descriptive analysis

A Report

L

9 12 2021

Synopsis

This paper offeres decision-makers to evaluate the outcomes of severe weather events in the USA. Two aspects are covered, namely which weather events causes major (1) harmful effects on population health in terms of injuries and deaths and (2) property and crop damage in terms of US.Dollar. Raw data is retrieved from the National Weather Service collecting comprehensive data since 2000's. Preprocessing, descriptively analysing and reporting was performed in R-Studio. Results show, by far, most people are hurt or killed by turnados and excessive heat. Thunderstorm wind, tornadoes and floods destroy mostly property and crops, almost 25 Billion US Dollar each in the end of the 2000's.

Introduction

Weather events can have a tremendous outcome on both population health in terms of injuries and deaths and on property, causing economically losses. Thus it is of great importance to prepare for severe weather events to prevent major consequences. As weather cannot be controlled and resources are scare, this paper aims to describe what events are most harmful to offer decision-makers objective and trustworthy reporting. Questions dealth with are:

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?

The whole process from retrieving data to preprocessing, analysing and summarizing is made publicily available. R-Studio was chosen to incorporate everything into one document making it reproducible by others.

Method

Accessing data and documentations

The National Weather Service of the USA reports on weather events since the 1950's. However, comprehensive data collection only started in the 1990's. The database of the National Centeres for Environmental Information is freely accessible. Information on this dataset and how it was collected were provided. All links were accessed on 9th of December, 2021.

Processing Data

For convenience, the dataset is stored in a zip-file. Packages to run this analysis were tidyverse and lubridate which needs to be loaded first.

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.5
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.3     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.0     v forcats 0.5.1
```

```
## Warning: package 'tibble' was built under R version 4.0.5
```

```
## Warning: package 'tidyr' was built under R version 4.0.5
```

```
## Warning: package 'readr' was built under R version 4.0.5
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
## Warning: package 'forcats' was built under R version 4.0.5
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 4.0.5
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 4.0.5
```

Loading the data can be made by the following, perhaps check `getwd()` first to know in which file the data will be loaded.

```
filename <- "repdata_data_StormData.csv"

if (!file.exists(filename)){
  download.file("https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2", filename)
  unzip(filename)
}

data <- read.table(filename, header = T, sep="," , na.strings = "NA")
```

As it is raw data, it's necessary to understand what classes are used.

```
str(data)
```

```
## 'data.frame':    902297 obs. of  37 variables:
##  $ STATE__    : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ BGN_DATE   : chr  "4/18/1950 0:00:00" "4/18/1950 0:00:00" "2/20/1951 0:00:00" "6/8/1951 0:00:00" .
##  $ BGN_TIME   : chr  "0130" "0145" "1600" "0900" ...
##  $ TIME_ZONE  : chr  "CST" "CST" "CST" "CST" ...
##  $ COUNTY     : num  97 3 57 89 43 77 9 123 125 57 ...
##  $ COUNTYNAME: chr  "MOBILE" "BALDWIN" "FAYETTE" "MADISON" ...
##  $ STATE      : chr  "AL" "AL" "AL" "AL" ...
##  $ EVTYPE     : chr  "TORNADO" "TORNADO" "TORNADO" "TORNADO" ...
##  $ BGN_RANGE  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ BGN_AZI    : chr  "" "" "" "" ...
##  $ BGN_LOCATI: chr  "" "" "" "" ...
##  $ END_DATE   : chr  "" "" "" "" ...
##  $ END_TIME   : chr  "" "" "" "" ...
##  $ COUNTY_END: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ COUNTYENDN: logi  NA NA NA NA NA NA ...
##  $ END_RANGE  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ END_AZI    : chr  "" "" "" "" ...
##  $ END_LOCATI: chr  "" "" "" "" ...
##  $ LENGTH     : num  14 2 0.1 0 0 1.5 1.5 0 3.3 2.3 ...
##  $ WIDTH      : num  100 150 123 100 150 177 33 33 100 100 ...
##  $ F          : int  3 2 2 2 2 2 2 1 3 3 ...
##  $ MAG        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ FATALITIES: num  0 0 0 0 0 0 0 0 1 0 ...
##  $ INJURIES   : num  15 0 2 2 2 6 1 0 14 0 ...
##  $ PROPDMG    : num  25 2.5 25 2.5 2.5 2.5 2.5 2.5 25 25 ...
##  $ PROPDMGEXP: chr  "K" "K" "K" "K" ...
##  $ CROPDMG    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ CROPDMGEXP: chr  "" "" "" "" ...
##  $ WFO        : chr  "" "" "" "" ...
##  $ STATEOFFIC: chr  "" "" "" "" ...
##  $ ZONENAMES  : chr  "" "" "" "" ...
##  $ LATITUDE   : num  3040 3042 3340 3458 3412 ...
##  $ LONGITUDE  : num  8812 8755 8742 8626 8642 ...
##  $ LATITUDE_E: num  3051 0 0 0 0 ...
##  $ LONGITUDE_: num  8806 0 0 0 0 ...
##  $ REMARKS    : chr  "" "" "" "" ...
##  $ REFNUM     : num  1 2 3 4 5 6 7 8 9 10 ...
```

As the data is historically stored and not comprehensive from 1950's it will be necessary to filter certain

years. However, to do this, first, the dates of beginning and ending of weather events needs to be transformed to class 'date'. Moreover, the year of beginning is substracted.

```
#transforming
data$BGN_DATE  <- as.Date(data$BGN_DATE, "%m/%d/%Y")
data$END_DATE <- as.Date(data$END_DATE, "%m/%d/%Y")

#mutating new columns
data <- data %>% mutate(Year = format(BGN_DATE, "%Y"))
```

The weather event is reported in the column `EVTYPE`. To answer question 1 columns `FATALITIES` and `INJURIES` are used. To answer question 2 `PROPDMG`, `PROPDMGEXP`, `CROPDMG` and `CROPDMGEXP` are used. The shortcut 'Prop' documents costs of damage to property whereas 'crop' documents the costs of damage to crops. Please note the digits within the EXP-columns. These signify the magnitude of damage caused in either K = thousand, B = Billion and M = Million. To cover this, new columns are mutated to make calculations easier.

```
data <- data %>%
    mutate(PROP_COSTS =
              case_when(
                PROPDMGEXP    == "K" ~ PROPDMG * 100000,
                 PROPDMGEXP   == "M" ~ PROPDMG * 1000000,
                   PROPDMGEXP  == "B" ~ PROPDMG * 1000000000,
                     TRUE ~ PROPDMG
              ),
           CROP_COSTS =
             case_when (CROPDMGEXP == "K" ~ CROPDMG * 100000,
                          CROPDMGEXP == "M" ~ CROPDMG * 1000000,
                           CROPDMGEXP == "B" ~ CROPDMG * 1000000000,
                            TRUE ~ CROPDMG
                        )

             )
```

Results

Population Health

As there are 900+ event types in this data set, this first result shows top 10 weather events impacting human health.

```
eventoccurence <- data %>% filter(Year >= 2000) %>%
        group_by(EVTYPE) %>%
        summarise(freq = n_distinct(BGN_DATE),
                  tot_inj = sum(INJURIES),
                  tot_fat = sum(FATALITIES),
                  tot_health = sum(FATALITIES) + sum(INJURIES)) %>%
        arrange(desc(tot_health))

healthevents5 <- head(eventoccurence$EVTYPE,5)

deaths <- eventoccurence[1,4]
injuries <- eventoccurence[1,3]
```

```
kable(head(eventoccurence,10))
```

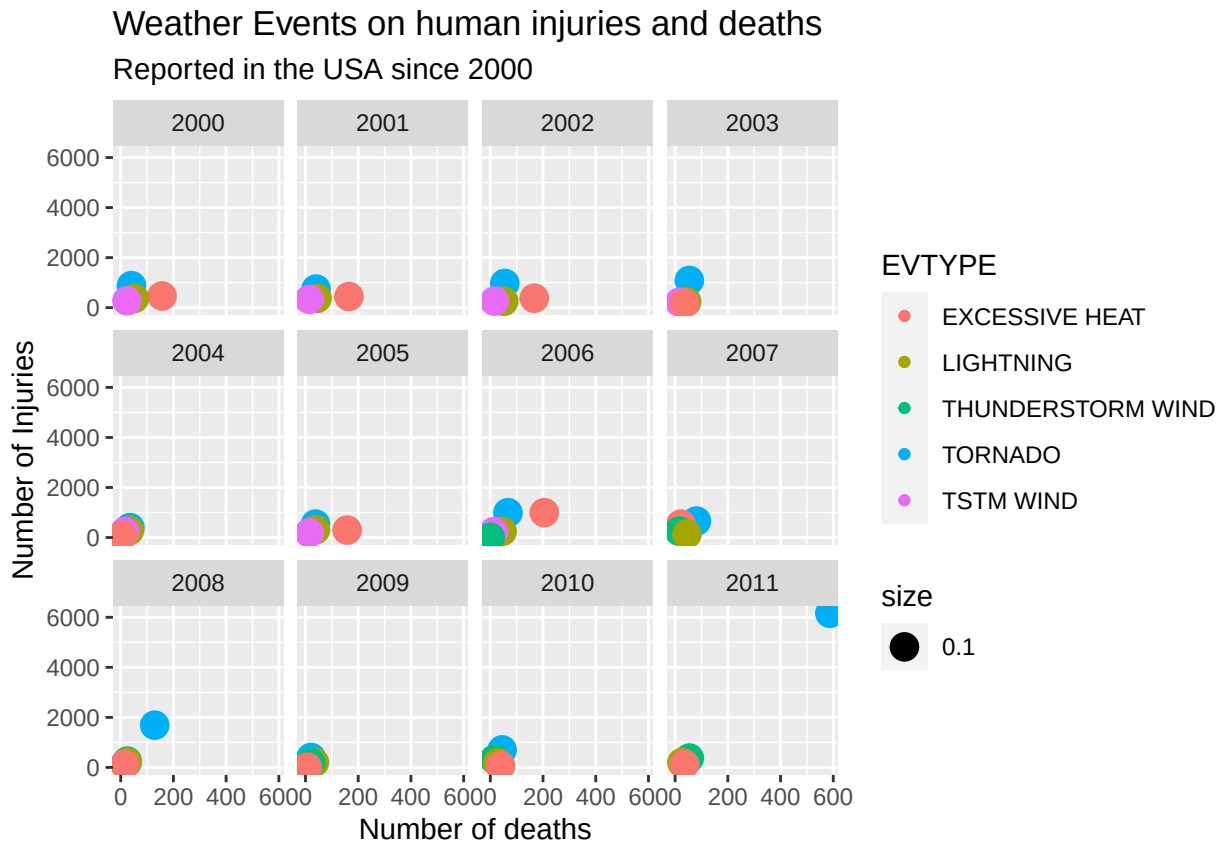
EVTYPE	freq	tot_inj	tot_fat	tot_health
TORNADO	2129	15213	1193	16406
EXCESSIVE HEAT	479	3708	1013	4721
LIGHTNING	2140	2993	466	3459
TSTM WIND	1662	1753	116	1869
THUNDERSTORM WIND	1294	1400	130	1530
HEAT	275	1222	231	1453
FLASH FLOOD	2740	812	600	1412
HURRICANE/TYPHOON	43	1275	64	1339
WILDFIRE	1243	911	75	986
HIGH WIND	2447	677	131	808

Most impact on population health ever since 2000 has the event type tornados, causing highest number of deaths (n = 1193) and by far greatest amount of injuries (n = 15213).

```
# aggregate per year
dataplot <- data %>% filter(Year >= 2000 & EVTYPE %in% healthevents5) %>%
  group_by(EVTYPE, Year) %>%
  summarise(freq = n_distinct(BGN_DATE),
            tot_inj = sum(INJURIES),
            tot_fat = sum(FATALITIES),
            tot_health = sum(FATALITIES) + sum(INJURIES)) %>%
  arrange(desc(tot_health))
```

```
## `summarise()` has grouped output by 'EVTYPE'. You can override using the `.groups` argument.
```

```
# show panel plot

 dataplot %>%
  ggplot(aes(x=tot_fat, y= tot_inj, color = EVTYPE, size= 0.1))+
  geom_point() +
  facet_wrap(~Year) +
  labs (x = "Number of deaths", y= "Number of Injuries",
        title="Weather Events on human injuries and deaths",
        subtitle = "Reported in the USA since 2000")
```



In this figure, it becomes clear that in the decade of 2000 tornados caused most injuries whereas excessive heat killed people at most. Thunderstorm wind, TSTM wind and lightning all share throughout this decade same number of injures and deaths as they all gather around <100 deaths and <1000 injuries. Only in the last year of data collection from this dataset, in 2011, a tornardo is by far top 1 severe weather event impacting population health.

Economical Costs due to damage

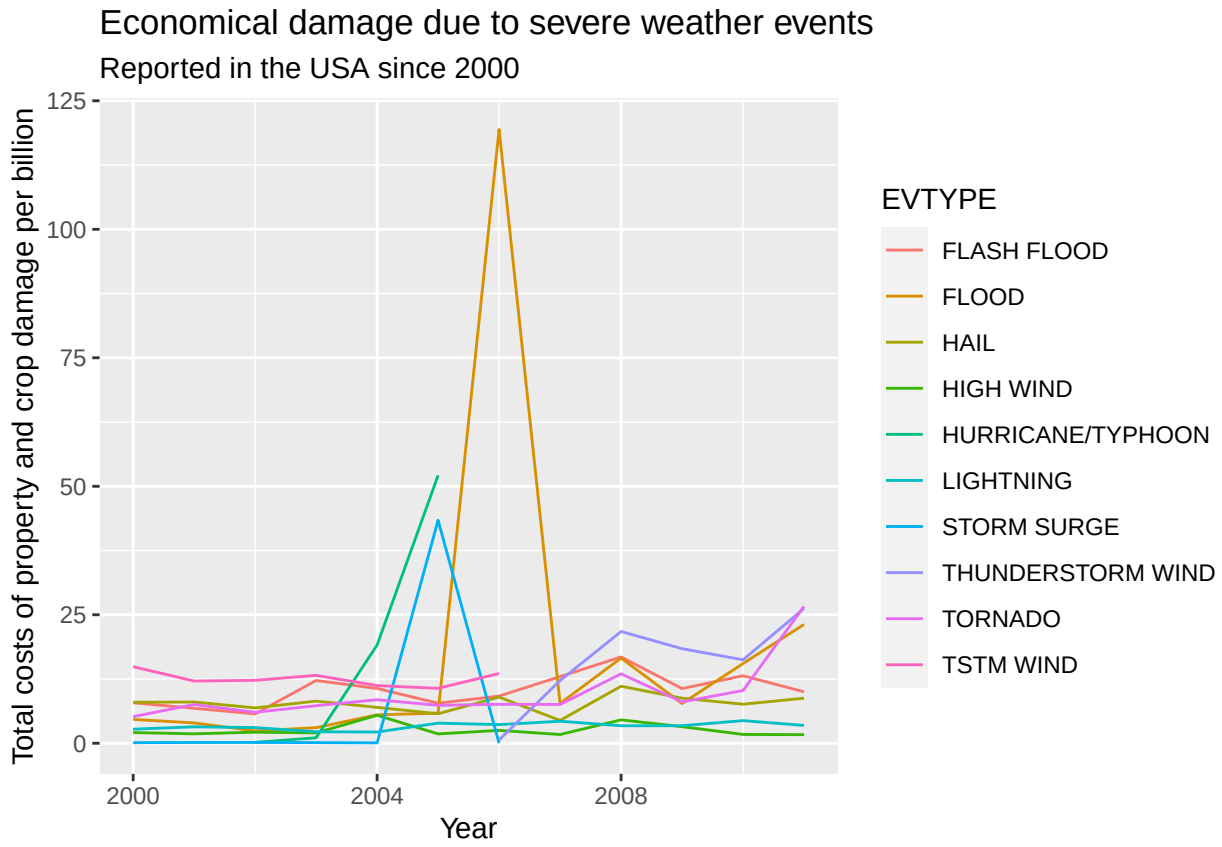
Same amount of possible events is there for economical burden thus following figure is limited to top 10 most costly severe weather events.

```
ecoevent <- data %>% filter(Year >= 2000) %>%
      group_by(EVTYPE) %>%
      summarise(freq = n_distinct(BGN_DATE),
                tot_prop = sum(PROP_COSTS),
                tot_crop = sum(CROP_COSTS),
                tot = sum(PROP_COSTS) + sum(CROP_COSTS)) %>%
      arrange( desc(tot))

ecoevents <- head(ecoevent$EVTYPE, 10)

dataplotcosts <- data %>% filter(Year >= 2000 & EVTYPE %in% ecoevents) %>%
  group_by(EVTYPE, Year) %>%
  summarise(freq = n_distinct(BGN_DATE),
            tot_prop = sum(PROP_COSTS),
            tot_crop = sum(CROP_COSTS),
            tot = (sum(PROP_COSTS) + sum(CROP_COSTS)) / 1000000000) %>%
```

```
  arrange( desc(tot))
```

```
## `summarise()` has grouped output by 'EVTYPE'. You can override using the `.groups` argument.
```

```
dataplotcosts %>%
  ggplot(aes(x=as.integer(Year), y= tot, color = EVTYPE))+
  geom_line() +
  labs (x = "Year", y= "Total costs of property and crop damage per billion",
        title="Economical damage due to severe weather events ",
        subtitle = "Reported in the USA since 2000")
```



In this figure it becomes apparent that in 2006 floods was in this whole decade major reasons for economical burden and outranges by far all other most costly weather events. In the following years, cost due floods remain constantly high. However, in the end of 2000's two types of winds cause major property and crop damage, namely thunderstorm wind and tornados.