ISYE_6501_HW3

Andrey Sivyakov

September 7, 2019

Question 5.1

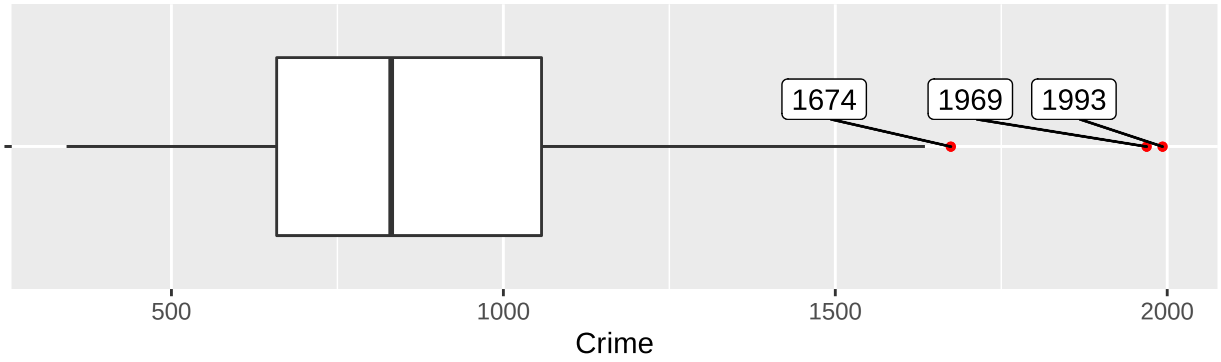
Using crime data from the file uscrime.txt(http://www.statsci.org/data/general/uscrime.txt, description at http://www.statsci.org/data/general/uscrime.html), test to see whether there are any outliers in the last column (number of crimes per 100,000 people). Use the grubbs.test function in the outliers package in R.

First, I will visualize the range of values in the Crime column using box plot.

```r
library(outliers)
library(car)
library(ggplot2)
library(ggrepel)

uscrime <- read.delim("uscrime.txt")
```

```r
outliers <- Boxplot(uscrime$Crime)
```

```r
out.labels <- rep("", nrow(uscrime))
out.labels[outliers] <- uscrime$Crime[outliers]

p <- ggplot(uscrime, aes("", Crime)) +
  geom_boxplot(outlier.colour = "red",
               outlier.shape = 16,
               outlier.size = 1.5) +
  coord_flip() +
  geom_label_repel(aes(label = out.labels),
                           hjust = 2,
                           direction = "x",
                           nudge_x = .2) +
  theme(axis.title.y = element_blank())
p
```



As we can see, three data points lie in the fourth quartile, but they are not located extremely far from the mean. Now let's apply Grubbs' test to see if the outlier with the largest difference from the mean (the observation equal 1993) is statistically different than the other values.

```
grubbs.test(uscrime$Crime, type = 10)
```

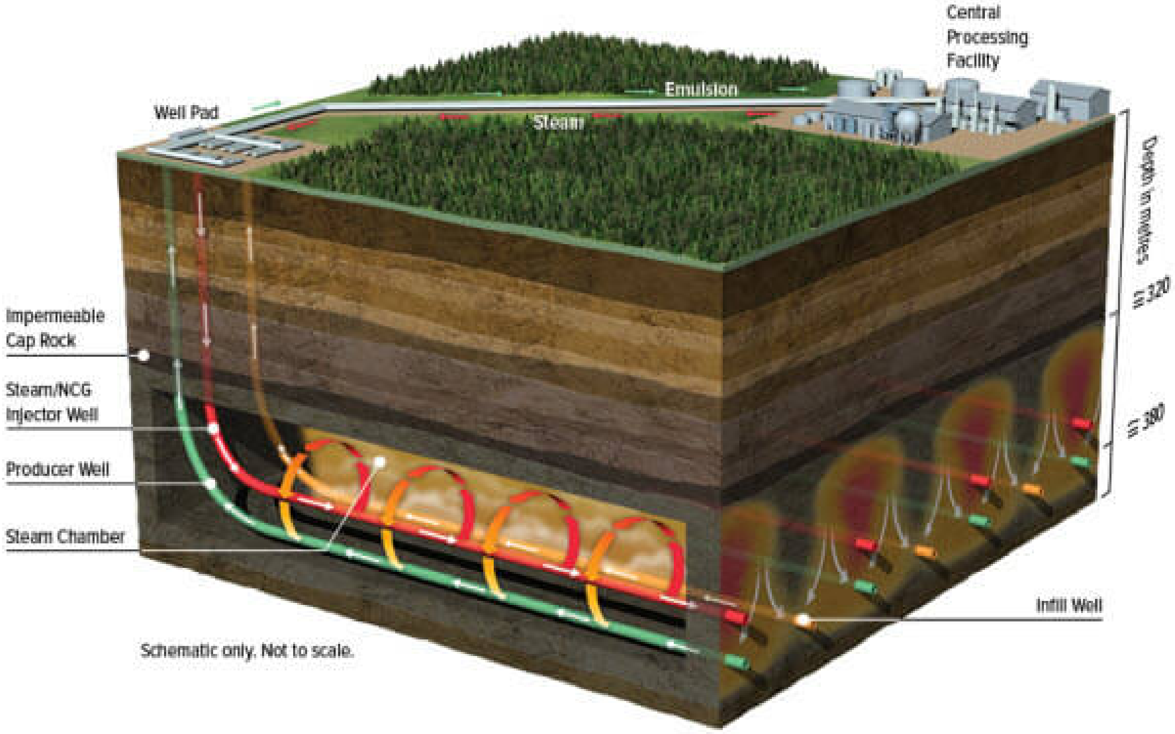
```
##
##  Grubbs test for one outlier
##
## data:  uscrime$Crime
## G = 2.81287, U = 0.82426, p-value = 0.07887
## alternative hypothesis: highest value 1993 is an outlier
```

According to the test results, G score is approximately 2.8, which means that this data point is 2.8 standard deviations away from the mean. P-value of the test is less than 0.05, thus, at 95% confidence level we can't reject the NULL hypothesis, so I conclude that the largest data point equal 1993 is not statistically different then the other values.

Question 6.1

Describe a situation or problem from your job, everyday life, current events, etc., for which a Change Detection model would be appropriate. Applying the CUSUM technique, how would you choose the critical value and the threshold?

Steam-assisted gravity drainage (SAGD; "Sag-D") is an enhanced oil recovery technology for producing heavy crude oil and bitumen. In SAGD operations, pairs of stacked horizontal wells are drilled into the reservoir about 400 metres beneath the surface. The top well injects steam to heat the bitumen, which separates from the sand and collects with the produced water in the lower well, approximately five metres below. The bitumen is then pumped to the surface, where it is separated from the water. The water is treated and recycled into the system (https://en.wikipedia.org/wiki/Steam-assisted_gravity_drainage).



The recycled water contains impurities, some of which settle down on the pipe walls, build up scum and eventually obstruct water flow in to the wells. Since concentration of harmful impurities in the recycled water can be measured, several techniques have been offered to predict the scum formation and suggest when scheduled maintenance of the water supplying system is needed. One of these techniques is analysing stationarity in concentration of impurities in recycled water over time. Lack of stationarity suggests that concentration of a certain substance is rapidly growing in the pipes, which will likely lead to a scum formation in near future.

I think CUSUM technique could be deployed to analyze concentration of harmful impurities in the recycled water and predicted costly failure of the SAGD system supplying recycled water. Critical values may be chosen based on lab simulation aimed at finding correlation between concentration of impurities in the recycled water and the speed of scum formation.

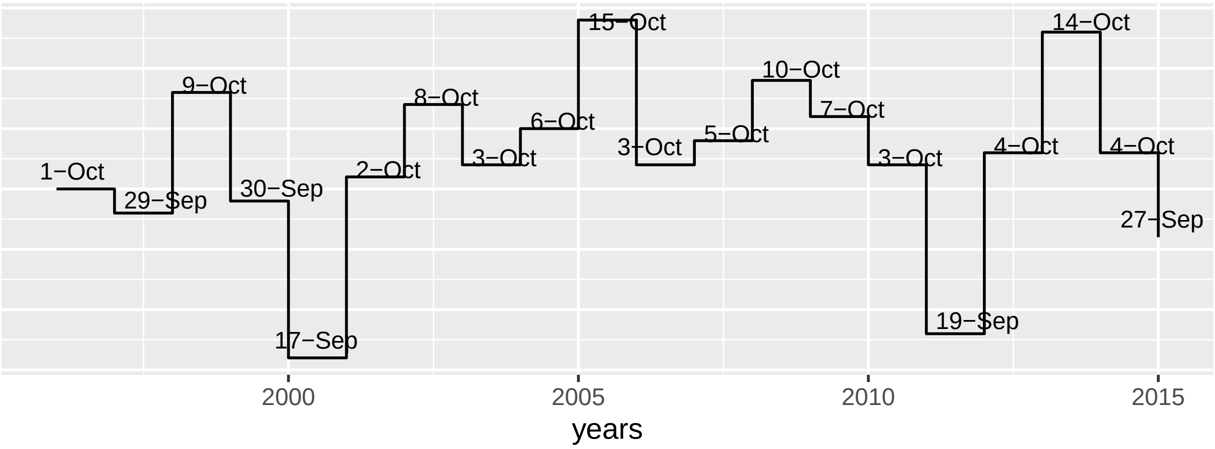
Question 6.2

1.Using July through October daily-high-temperature data for Atlanta for 1996 through 2015, use a CUSUM approach to identify when unofficial summer ends (i.e., when the weather starts cooling off) each year. You can get the data that you need from the file temps.txt or online, for example at http://www.iweathernet.com/atlanta-weather-records or https://www.wunderground.com/history/airport/KFTY/2015/7/1/CustomHistory.html. You can use R if you'd like, but it's straight forward enough that an Excel spreadsheet can easily do the job too.

To answer this question, I need to arbitrarily assign threshold (T) value for temperature observations in each year. I decided to use temperature value with 0.25 or less probabilty of being observed as a threshold. Thus, the date when St-statistic reaches the threshold level will represent unofficial summer end.

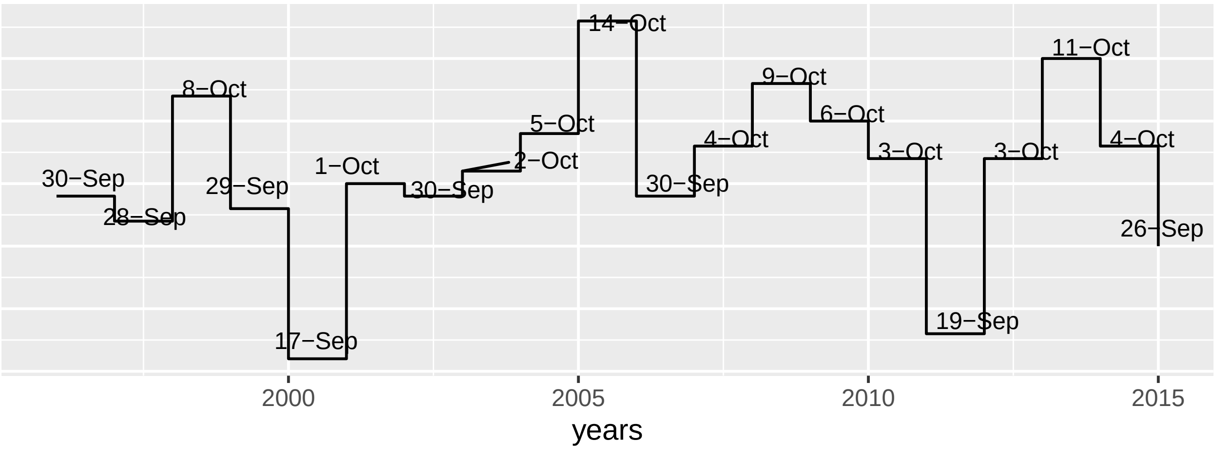
```r
temps <- read.delim("temps.txt", stringsAsFactors = F)

# user-defined function to calculate T-value for each year
def_t <- function(v) {
  qnorm(0.25, mean = mean(v), sd = sd(v))
}

# create empty vector to collect end-of-summer dates
out <- c()

for (col in 2:ncol(temps)) {
  t <- def_t(temps[, col])
  St <- c()
  for (i in 1:length(temps[, col])) {
    year_temps  <- temps[, col]
    mu <- mean(year_temps)
    if (i == 1) {
      St[i] <- max(0, (0 + mu - year_temps[i]))
    } else {
      St[i] <- max(0, (St[i-1] + mu - year_temps[i]))
    }
    if (St[i] >= t) {
      out[col-1] <- temps[i, "DAY"]
      break
    }
  }
}
```

Unofficial summer end in Atlanta, CUSUM, T = t[P(temp <= 0.25)], C = 0

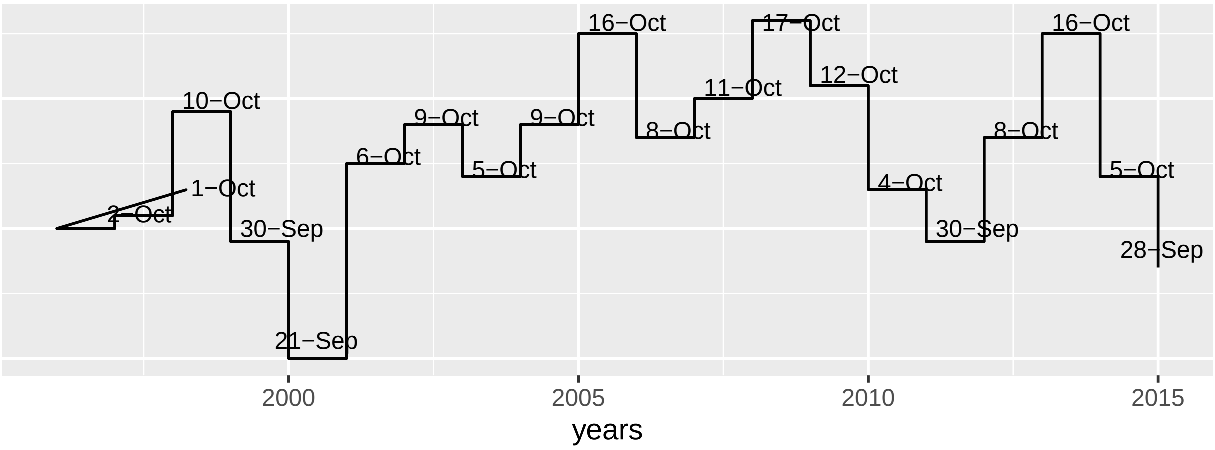


Using the same approach, I will try to identify unofficial summer end with different values of T and C

Unofficial summer end in Atlanta, CUSUM, T = t[P(temp <= 0.1)], C = 0



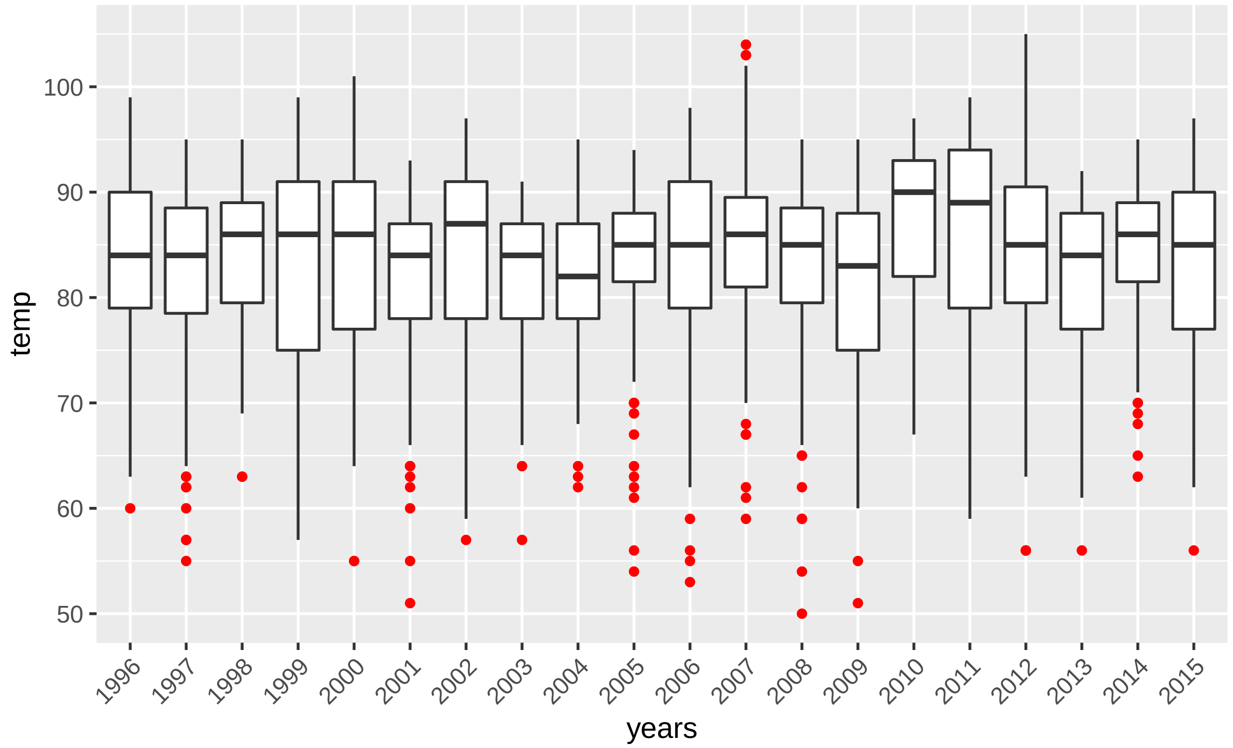
Unofficial summer end in Atlanta, CUSUM, T = t[P(temp <= 0.1)], C = 1



2.Use a CUSUM approach to make a judgment of whether Atlanta's summer climate has gotten warmer in that time (and if so, when).

According to the graphs in the previous answer, there is no obvious evidence that Atlanta's summer climate has gotten warmer in July-October within the observed period. Let's have a look at the distribution of the temperature each year.

```
temps <- read.delim("temps.txt", stringsAsFactors = F)

df <- data.frame(c(NULL), c(NULL))
years <- c(1996:2015)

for (i in 2:ncol(temps)) {
  temp <- cbind(rep(years[i-1], nrow(temps)), temps[, i])
  df <- rbind(df, temp)
}

names(df) <- c("years", "temp")
df$years <- as.factor(df$years)

ggplot(df, aes(years, temp)) +
  geom_boxplot(outlier.colour = "red",
               outlier.shape = 16,
               outlier.size = 1.5) +
  theme(axis.text.x = element_text(angle =45 , hjust = 1))
```



It looks like the median temperature stayed approximately stable, although seemingly higher variance was observed within the last 6-7 years. I conclude that there are no signs of Atlanta's summer climate getting warmer within the observed period.