ISYE 6501
Homework 3

September 10, 2019

Question 5.1

Before jumping into the Grubbs test, we performed a descriptive analysis of outliers using a boxplot.

```
# installing "outliers" and reading data
install.packages("outliers")
library(outliers)
data <- read.delim("uscrime.txt", header=TRUE)

# plotting data and requiring a bloxplot
plot(data$Crime)
boxplot <- boxplot(data$Crime)
boxplot$stats
boxplot$out
##[1] 1969 1674 1993
```

The bloxplot shows that the median of the data is 831, the 25th percentile 658.5, the 75th percentile 1057.5, and the extremes 342.0 and 1635. Observations 4 (Crime = 1969), 11 (Crime = 1674), and 26 (Crime = 1993) are detected as outliers. Therefore, the provisional expectation is that the Grubbs test will only identify outliers at the right end of the distribution.
We should be mindful that the Grubbs test can really only test for a) one outlier on one tail, b) two outliers on oppsite tails, and c) two outliers on one tail. The argument ”type” specifies either version of the test, and what we are looking for as a starter is b) because we want to test the presence of outliers at both ends. In this case, the Grubbs test-statistic tests the null-hypothesis that the two extreme entries are not jointly outliers, and specifically that either observation 27 (Crime = 342) OR observation 26 (Crime = 1993) IS NOT an outlier.

```
# requiring min and max before running the test
min <- which.min(data$Crime)
min
data[min,length(data)]
max <- which.max(data$Crime)
max
data[max,length(data)]

# requiring version b) of the Grubbs test
grubbs.test(data[,length(data)], type = 11)

##      Grubbs test for two opposite outliers
```

```
##data:  data[, length(data)]
##G = 4.26877, U = 0.78103, p-value = 1
##alternative hypothesis: 342 and 1993 are outliers
```

Clearly, with a p-value of 1 we accept the null-hypothesis which states that at least one between observation 27 and 26 is not an outlier. This does not exclude that outliers might be found at one tail of the distribution, and we suspect that the right tail is where this (or these) might be detected. Thus, we required a second Grubbs test specifying type = 10, the version of the test testing for one outlier on one tail.

```
grubbs.test(data[,length(data)], type = 10)

##      Grubbs test for one outlier

##data:  data[, length(data)]
##G = 2.81287, U = 0.82426, p-value = 0.07887
##alternative hypothesis: highest value 1993 is an outlier
```

The p-value associated with the test-statistic is more than 0.05 but only marginally so. Unfortunately, the grubbs.test command with type = 20 did not run, so we could not check whether the two rightmost entries observation 4 (Crime = 1969) and 26 (Crime = 1993) are jointly outliers. That being said, outlier detection is as much art as it is science, and we cannot ignore prior evidence we had from the boxplot when making the call. In conclusion, we tentatively remove both observations 4 and 26 from the data (despite a Grubbs test-statistic being only significant at the 10% level). Further inspection would be needed to understand the nature of these entries (bad data?) and make claims on why these figures are so high, but it falls beyond the scope of the current answer.

Question 6.1

A situation that would require a Change Detection is a respiratory monitor for hospital patients under critical care. A respiratory monitor measures the patient's breathing rate as well as heart rate. For the critical value, we would choose some change in the breathing/heart rate as a baseline for a stable patient (say, standard deviations of breathing/heart rates divided by 2). For the threshold, we would choose some standard deviations away (say 3 or 4) from the critical value for breathing/heart rate; for patients in critical care, we might tighten that standard deviation in order to catch more subtle changes.

Question 6.2

First approach

This question can be tackled many different ways, not only because the CUSUM method *per se* requires individual judgment to come up with the critical parameter (C) and the threshold (T), but also because the problem can be approached differently. We will propose two of the approaches we discussed and attempted collectively.
The first approach refers to the excel file *cusum_a.xlsx*. The first spreadsheet ”DATA 1996-2015” reports daily temperatures and cumulative sum for each year. The formula used in the S_t column subtracts out the μ for each year (reported at the bottom) and sets the parameter C to 0. We then conditionally-formatted each S_t column to red-highlight the maximum value. It can be noted that the max was attained anywhere between August 31st and October 4th but we threw off this last one case – corresponding to year 2005 – as we deeemed it to be an outlier. In fact, it would represent the only time that summer ends in October, more than a week apart from the second latest date of September 25th in 2010. In the helper spreadsheet ”END OF SUMMER 1996-2015”, we indexed dates August 31st through September 25th – the min and max end

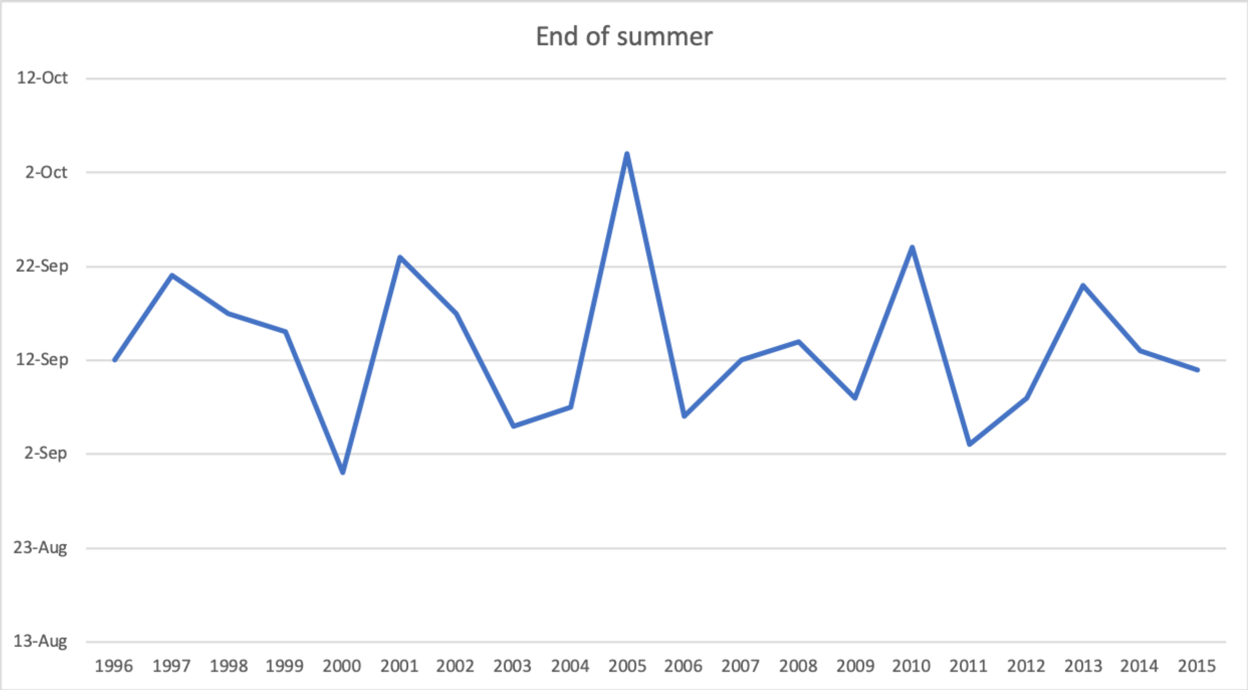

Figure 1: Trend 1996-2015

of summer – with the values 1 to 26 and computed the μ (13.4) and SD (6.8). According to this method, summer ended somewhere between September 12th and September 13th, on average.

What we did so far is preparatory to the last step shown in the spreadsheet ”FORECAST 2019”. Averages of day temperatures were taken across years, and entered in the x_μ column. We took the average of averages (83.3) and subtracted this out from the day averages along with the parameter C which is set to be one-half of the previously calculated SD (3.4). The last column applies CUSUM on the day averages and spits out our best prediction for the end of summer 2019, September 15th.

We immediately notice that this date is not much later than the previously computed mean end of summer (less of a quarter of a SD). Once plotting the CUSUM results from the spreadsheet ”END OF SUMMER 1996-2015_PLOT”, we obtain further support for the claim that onset of Fall has not been appreciably delayed over the years. There really seems to be no upward trend.

Second approach

The second approach (see file *cusum_b.xlsx*) follows a more straightforward implementation of the CUSUM method which guesses on the C and T parameters. Again, we took day temperature averages across years, and then found the μ and SD of temperatures for the months of July and August, which we deemed to be ”summer months”. For the critical value C and the threshold T, we tweaked the ’factor’ for each to 0.5 and 20 (that is, C = SD/0.5 and T = SD*20). Since we derived a T value of 21, we detected a change at September 14, when the S_t value reaches 22.72. September 14 is therefore our best guess for the end of summer 2019.

We replicated the CUSUM method (same formula, same C and T) for each year in the dataset and obtained again a trend which is to the very least erratic. Therefore, there really seems to be no straightforward evidence of an upsurge in average temperatures and delayed onset of the Autumn.

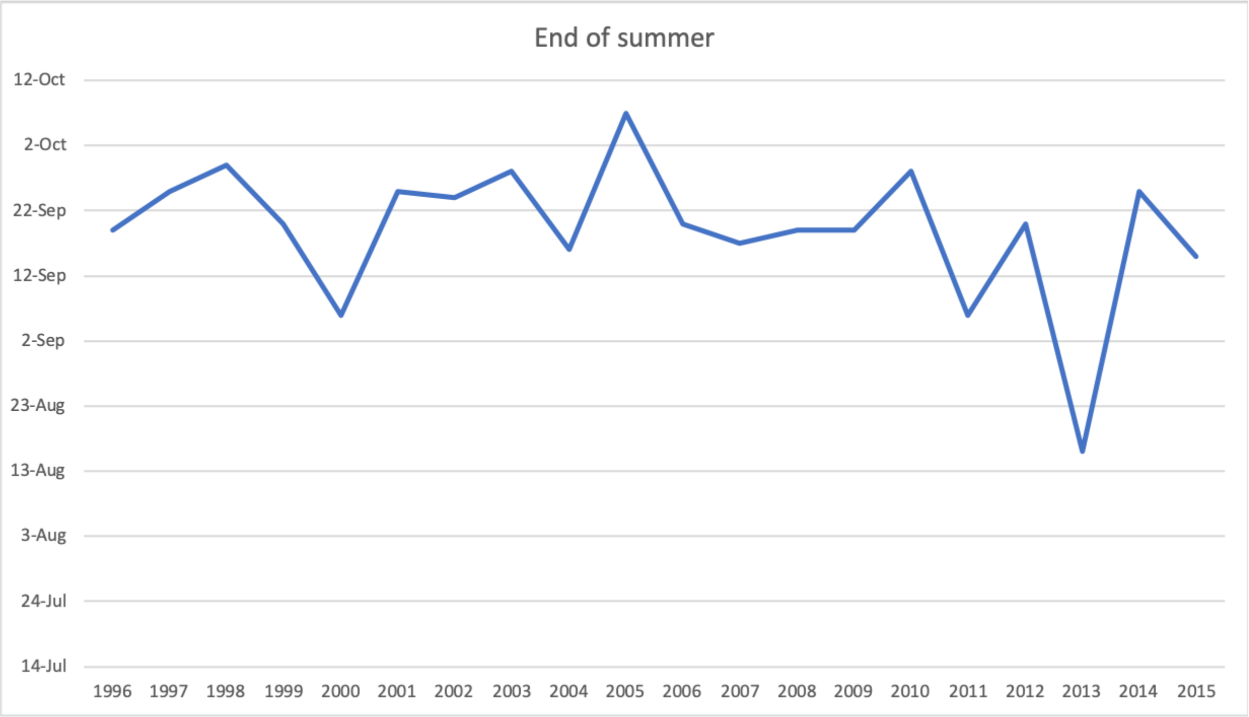

Figure 2: Trend 1996-2015