HW 3

Code

Question #1: Outliers

Loading up the data

Hide

```
require(outliers)
```

```
Loading required package: outliers
```

Hide

```
url <- "https://d37djvu3ytnwxt.cloudfront.net/assets/courseware/v1/17
b85cea5d0e613bf08025ca2907b62f/asset-v1:GTx+ISYE6501x+3T2017+type@ass
et+block/uscrime.txt"
df <- read.table(url, header = TRUE)
dim(df)
```

```
[1] 47 16
```

Hide

```
# Seperating the last column we will examine into a variable
crime_per_capita <- df[,'Crime']
length(crime_per_capita)
```

```
[1] 47
```

Hide

```
crime_per_capita
```

```
 [1]  791 1635  578 1969 1234  682  963 1555  856  705 1674  849  511
664  798  946  539  929
[19]  750 1225  742  439 1216  968  523 1993  342 1216 1043  696  373
754 1072  923  653 1272
[37]  831  566  826 1151  880  542  823 1030  455  508  849
```

Perfroming Grubbs test for outliers

To determine whether the highest and lowest values are potential outliers, I used the grubbs.test() function with type=10 which looks for one outlier. The first test looks at the highest value in the set and the second looks for the lowest value, indicated by using opposite=True.

Hide

```
# Testing for the highest value
grubbs.test(crime_per_capita, type=10, opposite = FALSE)
```

```
    Grubbs test for one outlier

data:  crime_per_capita
G = 2.81290, U = 0.82426, p-value = 0.07887
alternative hypothesis: highest value 1993 is an outlier
```

Hide

```
# Testing for the highest value
grubbs.test(crime_per_capita, type=10, opposite = TRUE)
```

```
    Grubbs test for one outlier

data:  crime_per_capita
G = 1.45590, U = 0.95292, p-value = 1
alternative hypothesis: lowest value 342 is an outlier
```

The returned results seem a bit curious to me. While the highest value (1993) is borderline ‘signigicant’ with a p value close to .05, the lowest value (342) is not remotely suspect as it has a p value of 1. I decided to look at the data a bit more closesly and check out some of the higher and lower values of the data.

Hide

```
crime_per_capita[crime_per_capita <500]
```

```
[1] 439 342 373 455
```

Hide

```
crime_per_capita[crime_per_capita >1300]
```

```
[1] 1635 1969 1555 1674 1993
```

The two values at 1900+ seem very high as compared to the rest of the data and the four lowest data points are all at roughly 350 +- 100.

After performing the grubbs test and looking at these data points I would not want to investigate the two 1900+ values further before deciding if they were true outliers or not. It's possible they are but also could be natural occurences. The low values could be coming from states with less crime and population, say Wyoming, while the two highest could come from areas with a propensity for above average crime rates (e.g. Illinois from Chicago)

Question #2: Change Detection Application

In the insurance world, cusum could be applied to customer claims. These could be done either in claim counts (the number of claims) or claim cost (dollar amount of claims). For something like auto insurance, the cusum could be applied to claim cost per customer per year. The threshold could be determined by profit margin of a customer. Taking the profit made on each customer and subtracting the average claim cost over a 12 year period could be a rule of thumb critical value. This threshold would be a rule of thumb as anytime it was passed would indicate the insurance company is no longer making money on that customer. As for a critical value, this would depend on the company's goals. If they were very concerned with making money on each customer, they would use a higher critical value (1.5) while if they were less concerned could use a lower one (.5)

Question #3.1: CUSUM for Summer's End This question focuses on finding when summer ends by way of examining daily temperature highs.

The cusum keeps a running total of deviation from the sample mean. In this case that sample mean is the average temperature during that summer period. As we are trying to find the point at which temperatures are cooling off (i.e. lower than the avg summer temperature from that point onward).

The R package qcc is a statistical control package that has a cusum function within it. Passing each year as the data for the cusum returns a cusum object and plot.

For mu, I considered using the average summer temperature across all years but ultimately decided to use the average summer temperature within a given year. Since the goal was to find the point at which temperatures start to decrease I thought it appropriate to use the average temperature of that summer to determine whether temperatures were decreasing.

I experimented with various threshold values, which within the cusum function are measured in standard errors. I ran various thresholds and ultimately decided on 7 SE to use as the threshold. The default is 5 so this slightly higher threshold would mean a change that registers is more likely to be a true change than a false alarm. I opted to

slightly increase the critical value from a default of 1 to 1.2 as this seemed to give better results for indicating a clear change from temperatures increasing or decreasing. With the slightly higher threshold but also slighlty increased sensitivity the results were promising.

The following code runs a loop producing a cusum plot for each year as well as a plot with the last day of summer as indicated by the cusum function: the first day at which a decreasing change is noted.

Hide

```
require(qcc)
```

```
Loading required package: qcc
Package 'qcc', version 2.6
Type 'citation("qcc")' for citing this R package in publications.
```

Hide

```
# Loading the data
url <- "https://d37djvu3ytnwxt.cloudfront.net/assets/courseware/v1/59
2f3be3e90d2bdfe6a69f62374a1250/asset-v1:GTx+ISYE6501x+3T2017+type@ass
et+block/temps.txt"
temps <- read.table(url, header = TRUE)
# Finding the average temperature for each summer to use as the mu
avg_summer_temp <- c()
for (i in 2:21){
  avg_summer_temp <- append(avg_summer_temp, mean(temps[,i]))
}
length(avg_summer_temp)
```

```
[1] 20
```

Hide

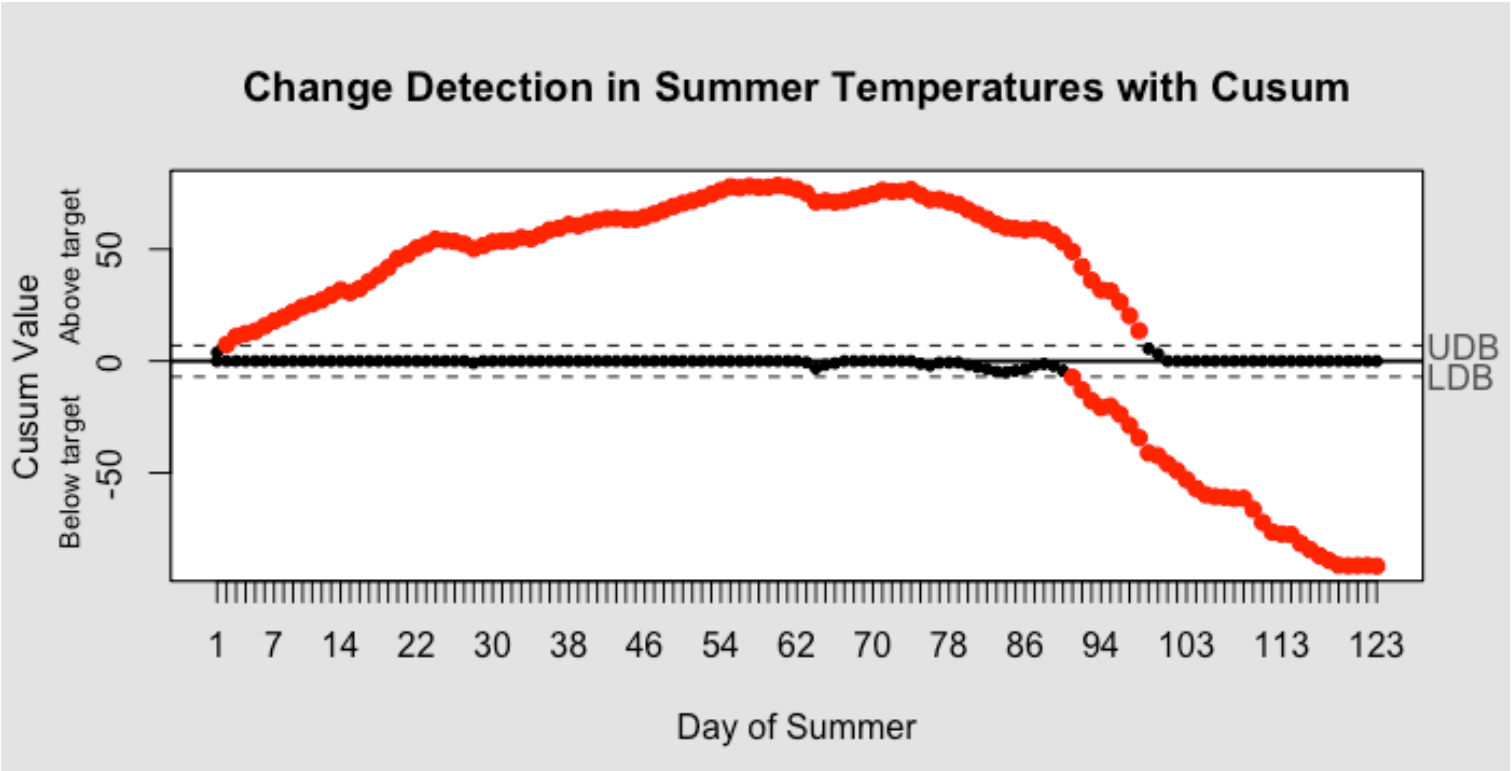
```
print(avg_summer_temp)
```

```
 [1] 83.71545 81.67480 84.26016 83.35772 84.03252 81.55285 83.58537 8
1.47967 81.76423 83.35772
[11] 83.04878 85.39837 82.51220 80.99187 87.21138 85.27642 84.65041 8
1.66667 83.94309 83.30081
```

Hide

```
# The average of the average summer temperatures is 83.33902
```

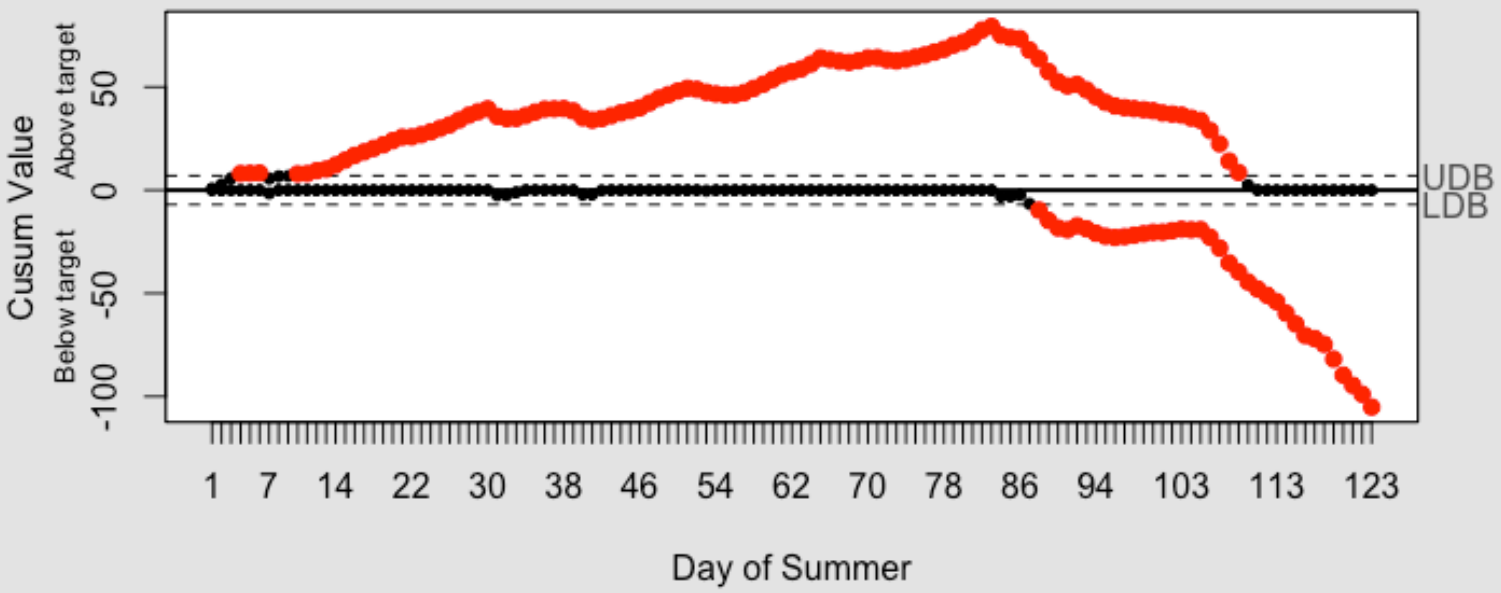
```
avg_avg_summer_temp <- mean(avg_summer_temp)
last_day_of_summer <- c()
for (year in 2:21){
  # Using each year's temps for cusum
  years_cusum <- cusum(temps[,year],
  # The center is the average of the temperatures. The below value is
the average across
  # all years. If it isn't used, the cusum function takes that summer
's average temperature
  #center = 83.33902
  # Decision interval is the threshold for when a change is detected

  # Will use 7
  decision.interval = 7,

  # The critical value (our senstivity to change). Will use 1.2
  se.shift = 1.2,

  # Chart name
  data.name = "Summer Temperatures",
  title = "Change Detection in Summer Temperatures with Cusum",
  xlab = "Day of Summer",
  ylab = "Cusum Value")

  # Second summer day where decreasing temps have passed decsion thre
shold
  # Chose the second day to avoid a fluke change detected
  last_day_of_summer <- append(last_day_of_summer, years_cusum$violat
ions$lower[2])
}
```



Number of groups = 123
Center = 83.71545
StdDev = 3.204569

Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 130

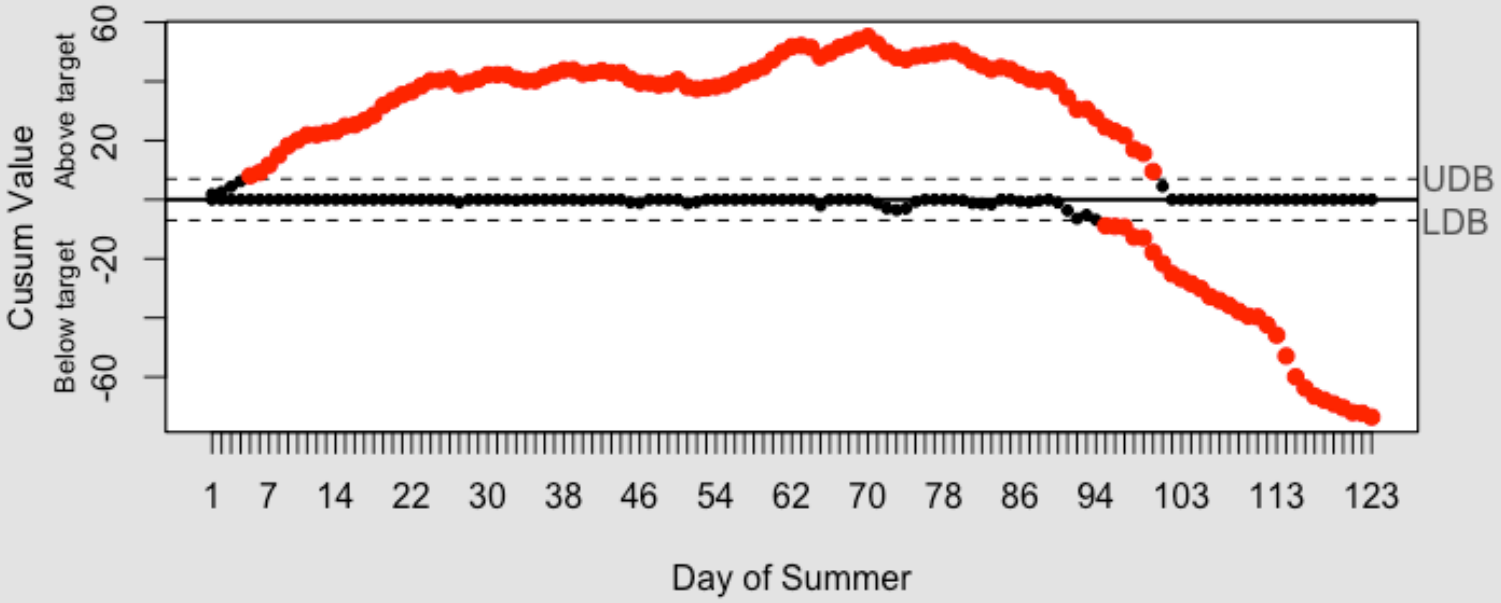
Change Detection in Summer Temperatures with Cusum



Number of groups = 123
Center = 81.6748
StdDev = 3.182769

Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 139

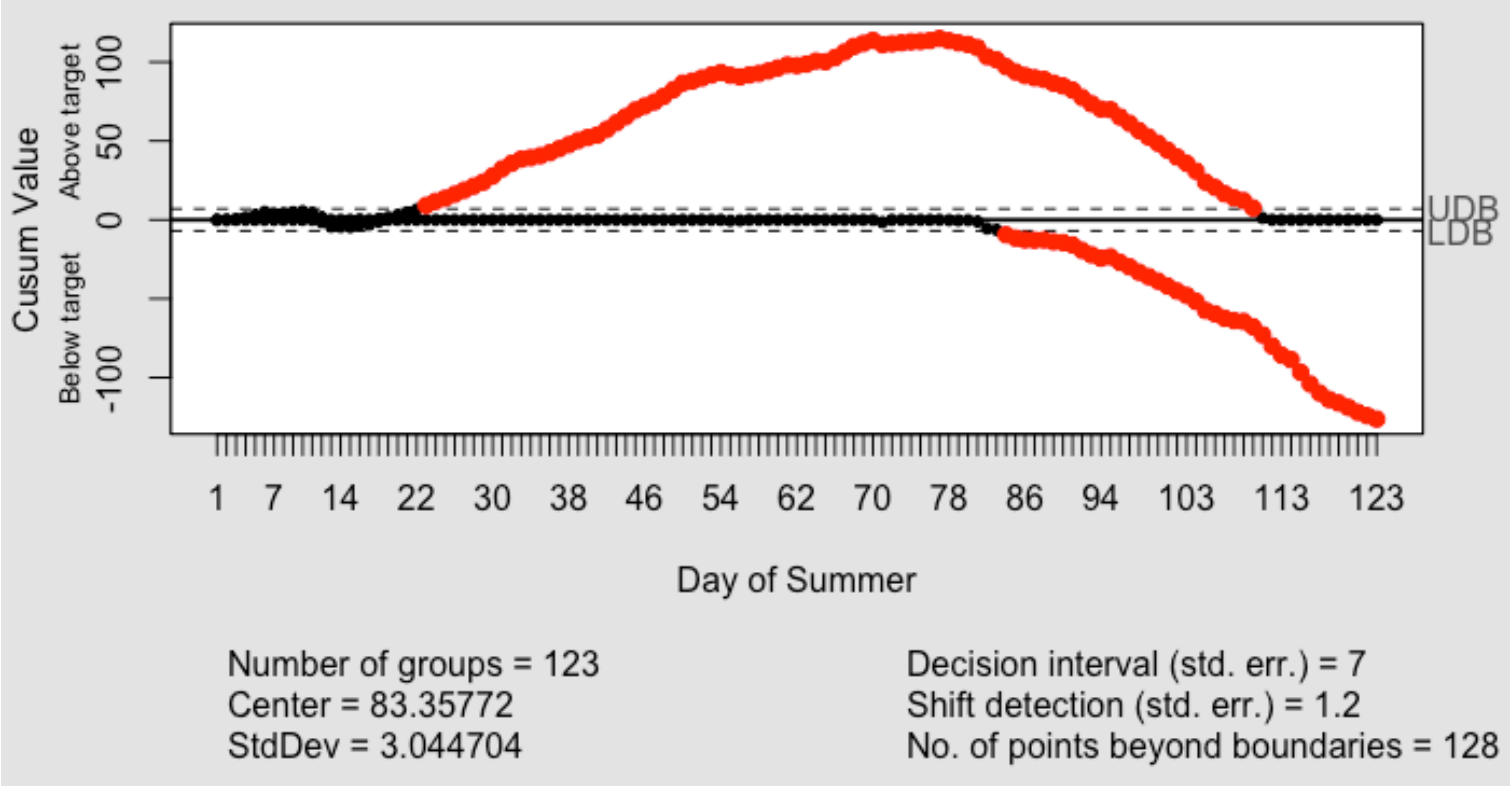
Change Detection in Summer Temperatures with Cusum



Number of groups = 123
Center = 84.26016
StdDev = 2.77584

Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 125

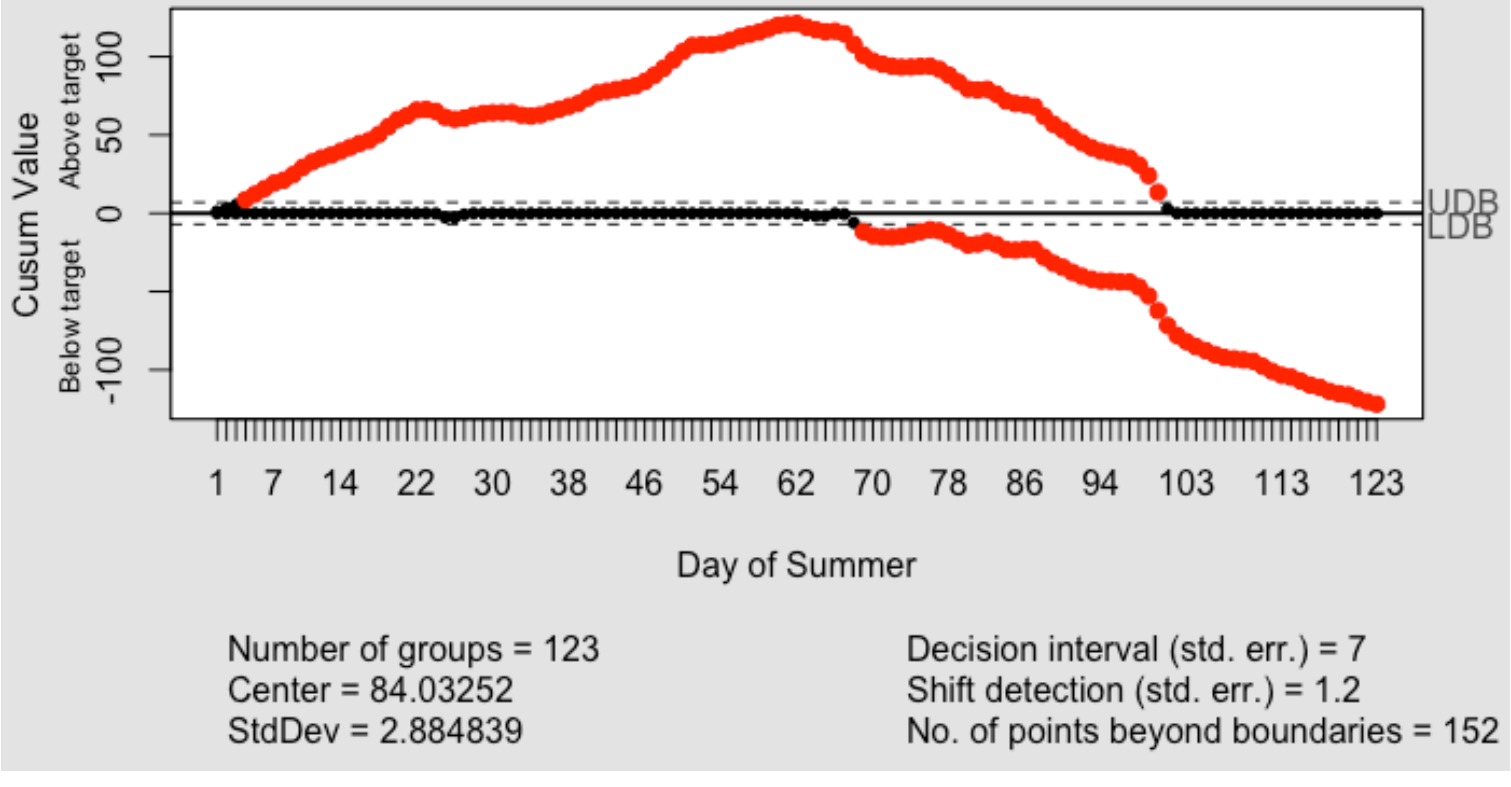
Change Detection in Summer Temperatures with Cusum



Number of groups = 123
Center = 83.35772
StdDev = 3.044704

Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 128

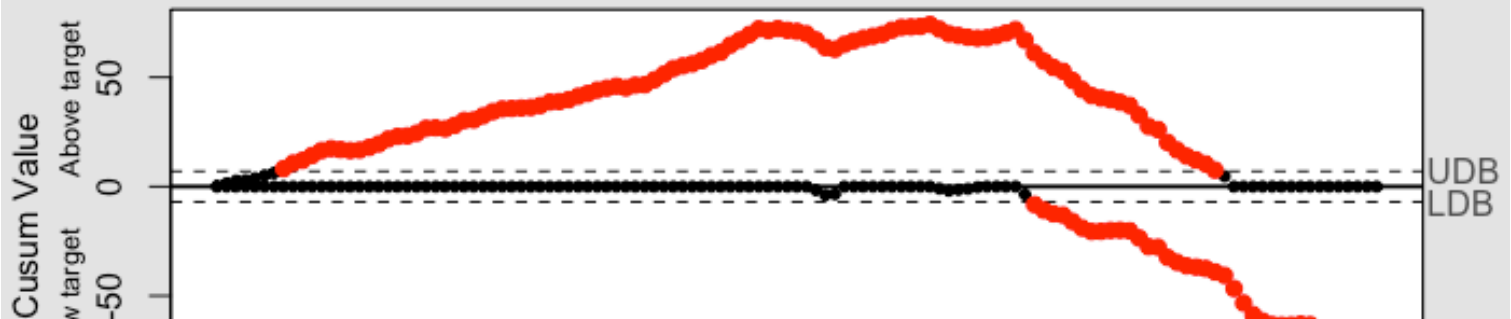
Change Detection in Summer Temperatures with Cusum



Number of groups = 123
Center = 84.03252
StdDev = 2.884839

Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 152

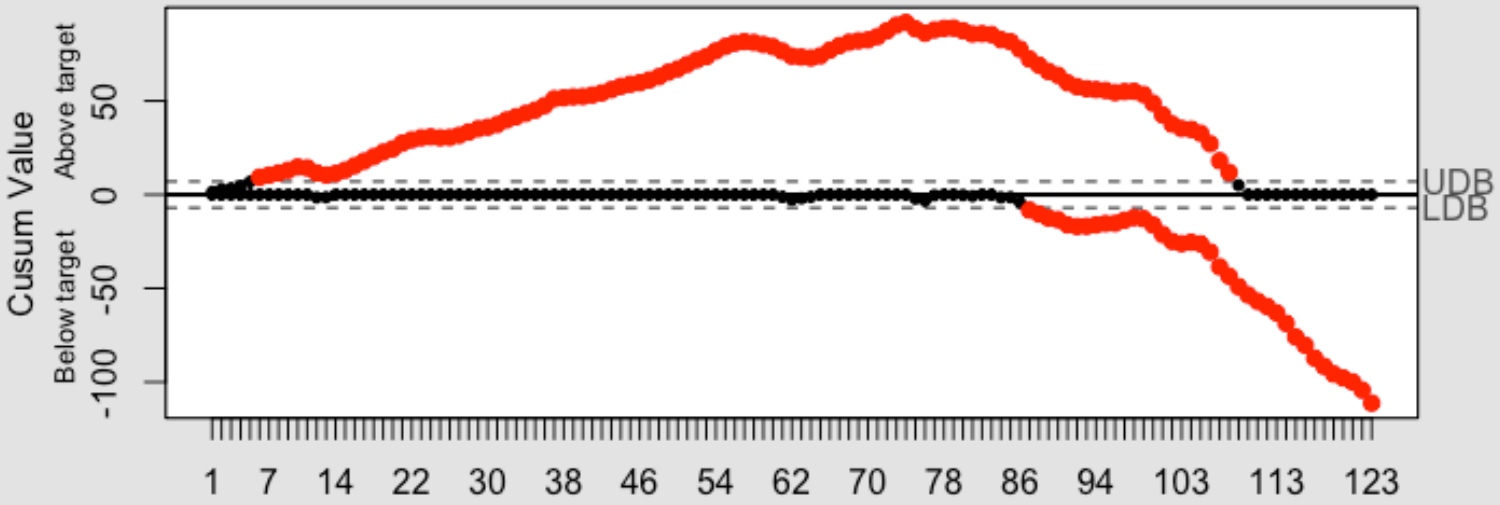
Change Detection in Summer Temperatures with Cusum



Number of groups = 123
Center = 81.55285
StdDev = 2.972038

Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 136

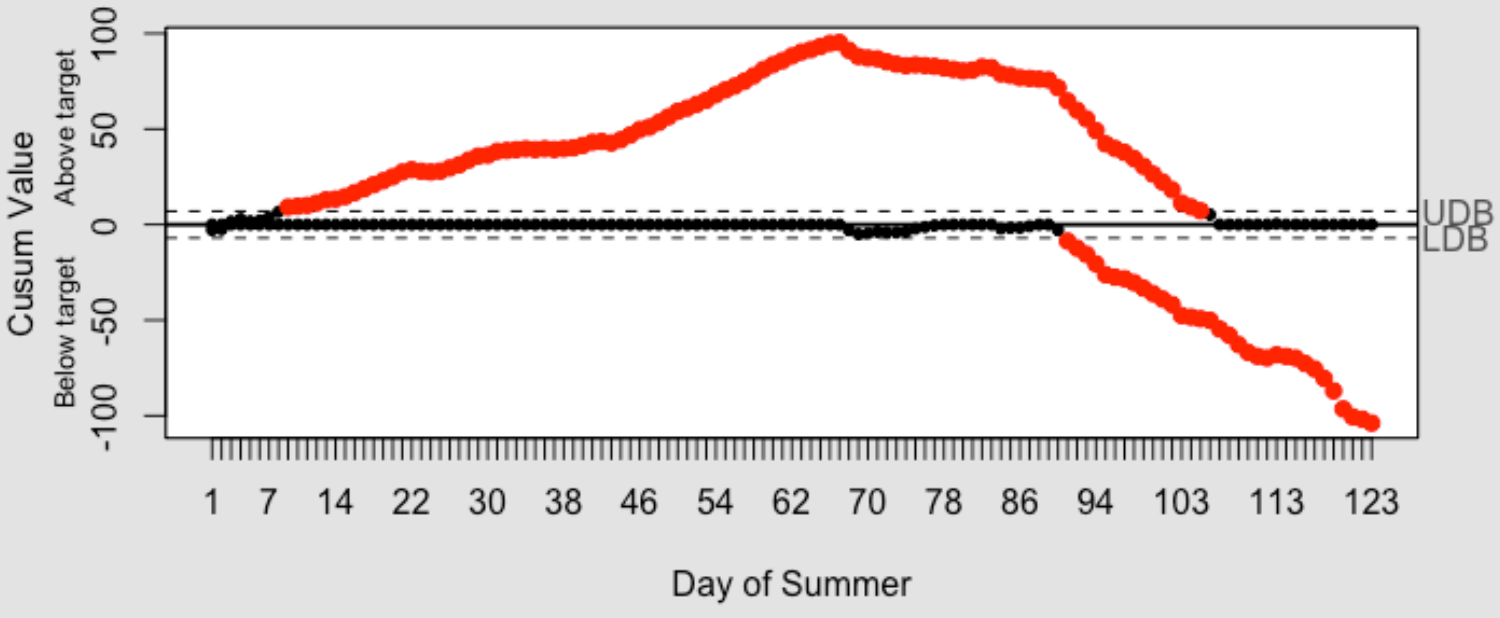
Change Detection in Summer Temperatures with Cusum



Number of groups = 123
Center = 83.58537
StdDev = 3.13917

Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 140

Change Detection in Summer Temperatures with Cusum

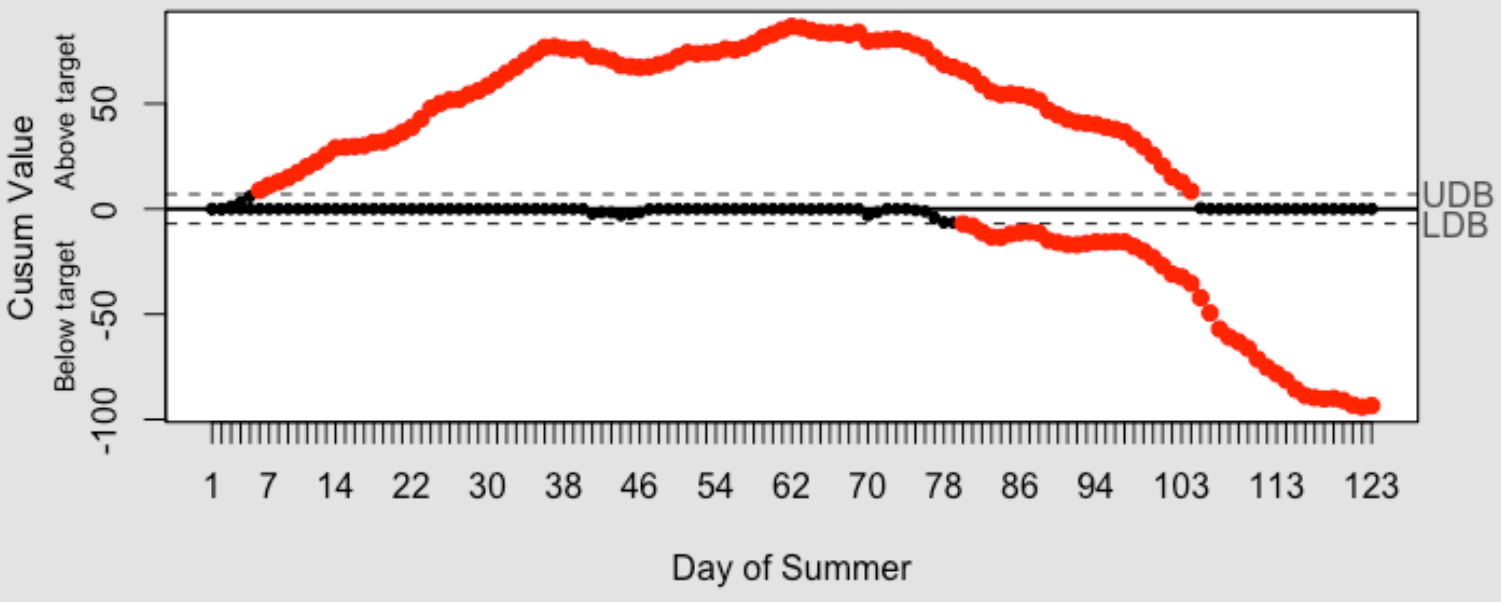


Number of groups = 123
Decision interval (std. err.) = 7

Number of groups = 123
Center = 81.47967
StdDev = 2.441577

Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 130

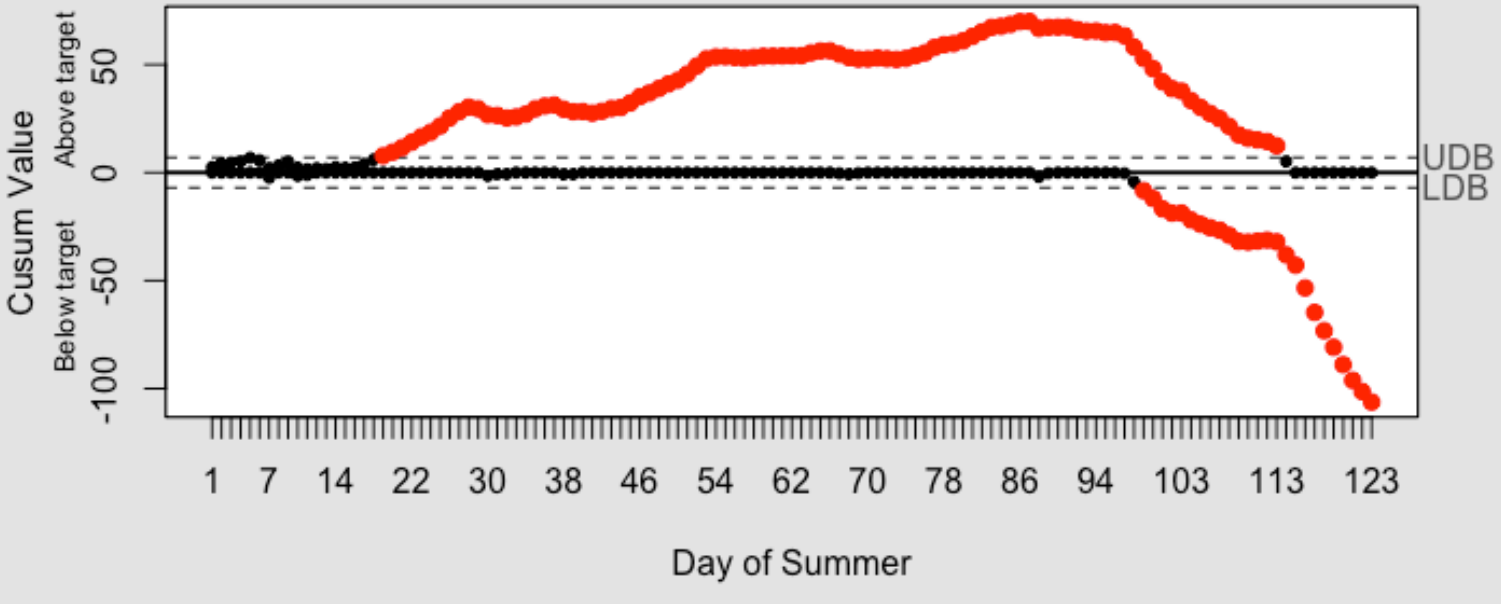
Change Detection in Summer Temperatures with Cusum



Number of groups = 123
Center = 81.76423
StdDev = 2.397977

Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 143

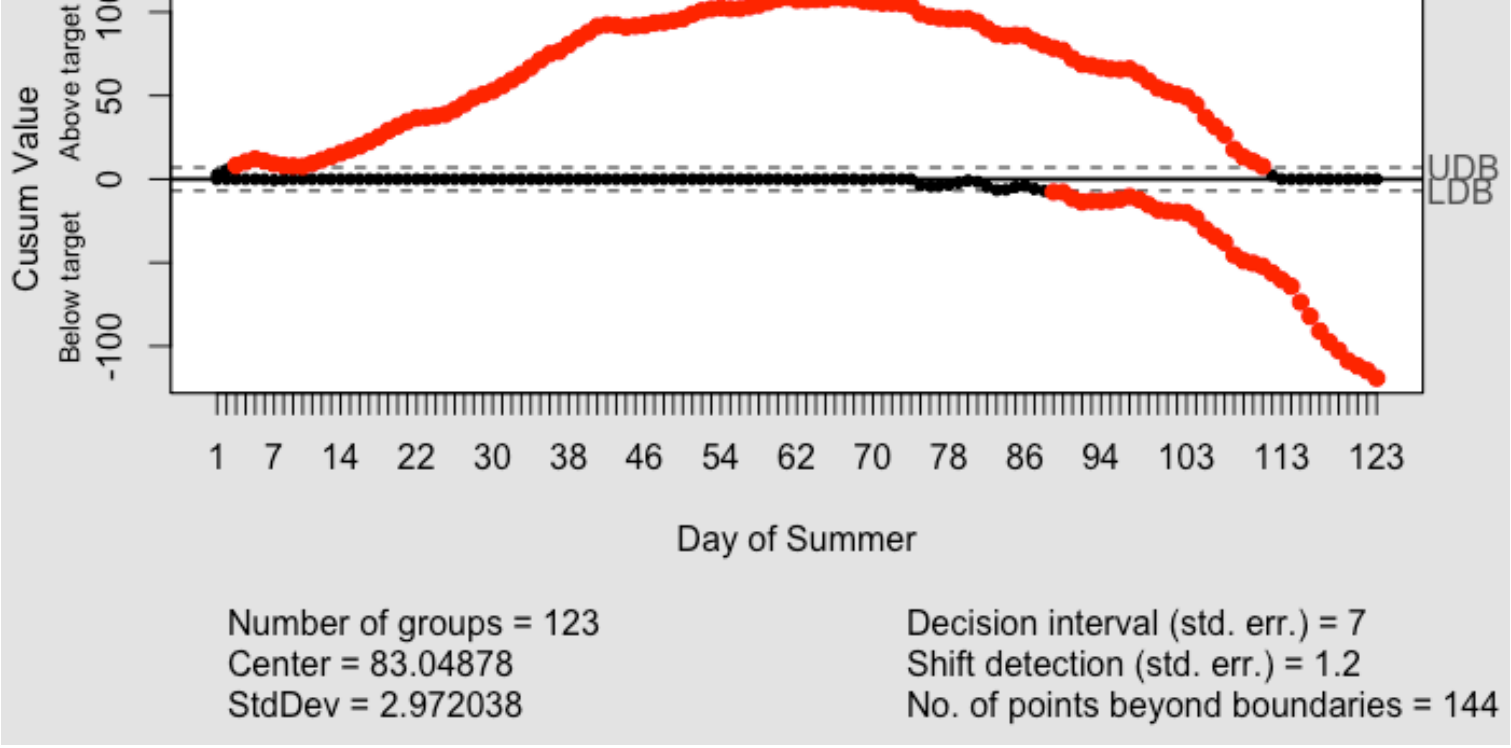
Change Detection in Summer Temperatures with Cusum



Number of groups = 123
Center = 83.35772
StdDev = 2.463376

Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 120

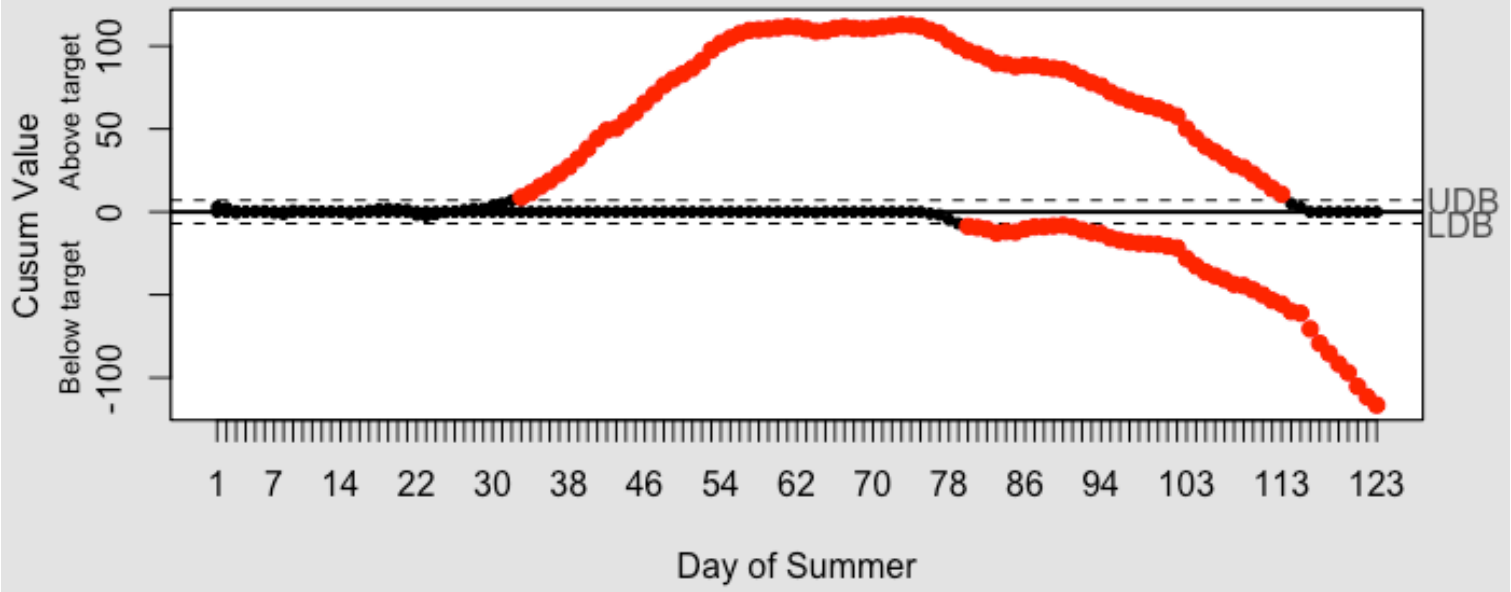
Change Detection in Summer Temperatures with Cusum



Number of groups = 123
Center = 83.04878
StdDev = 2.972038

Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 144

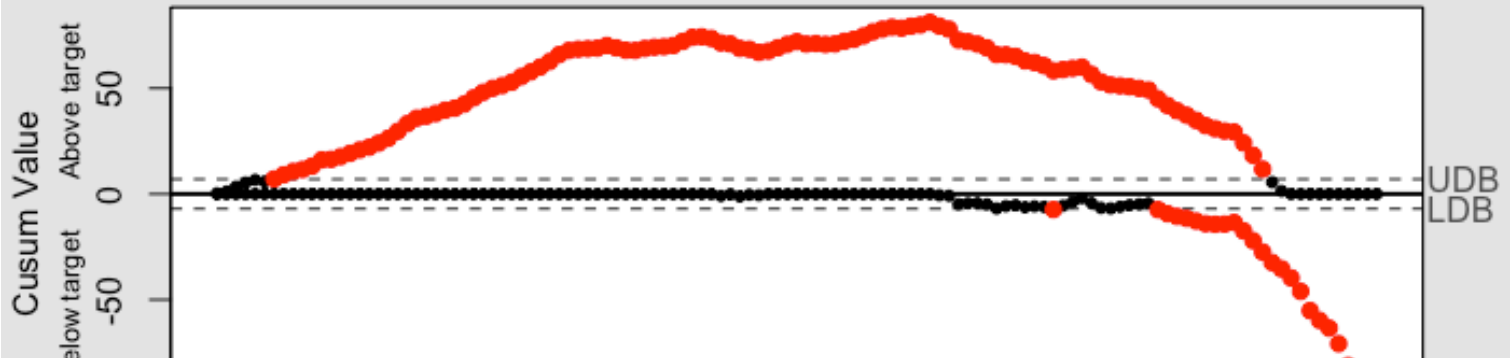
Change Detection in Summer Temperatures with Cusum

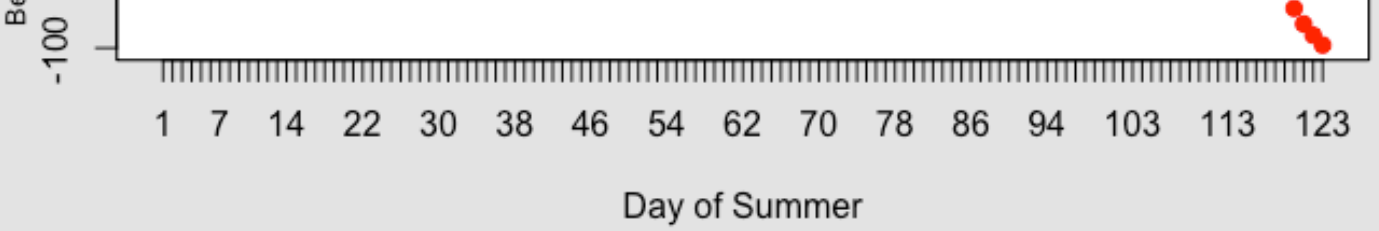


Number of groups = 123
Center = 85.39837
StdDev = 2.630508

Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 125

Change Detection in Summer Temperatures with Cusum

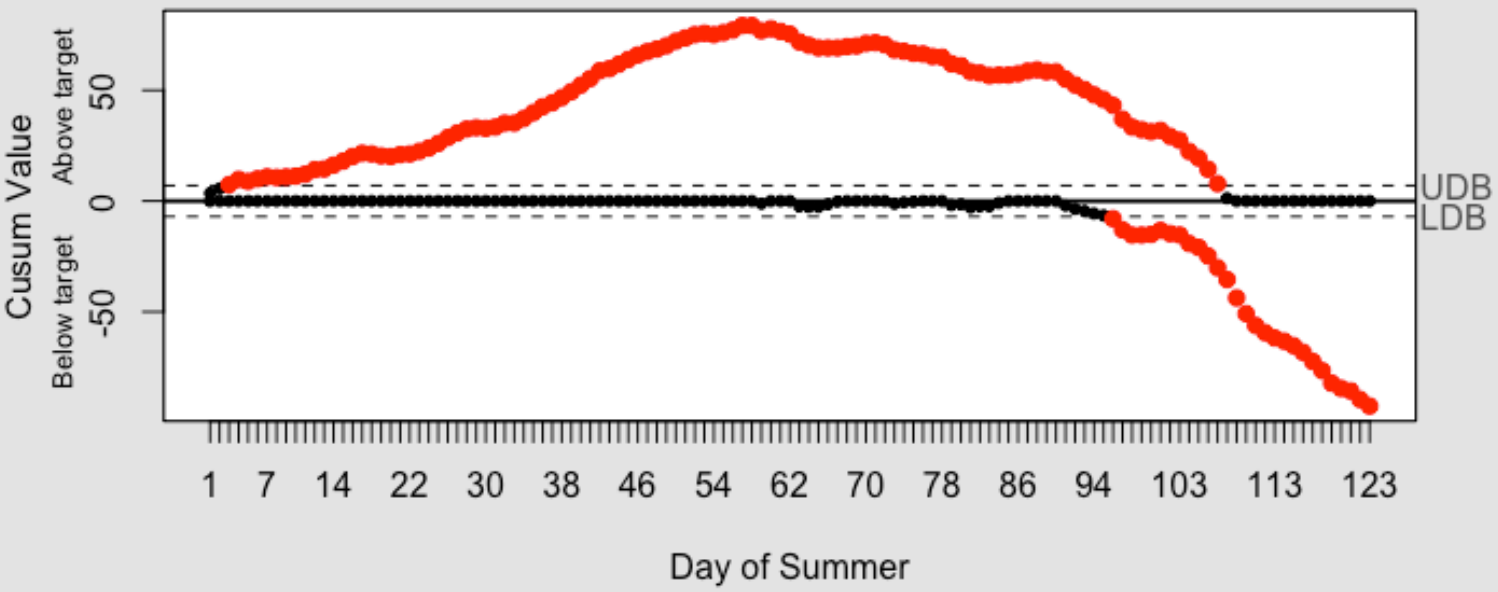




Number of groups = 123
Center = 82.5122
StdDev = 2.950238

Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 130

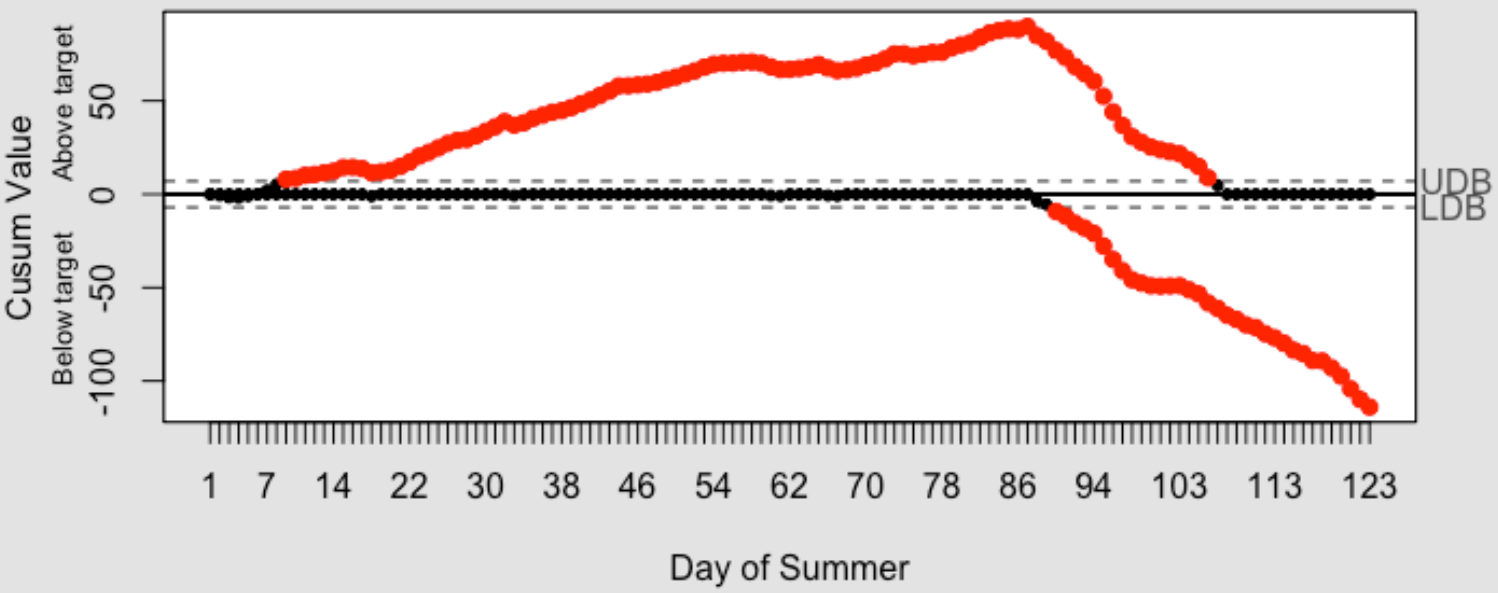
Change Detection in Summer Temperatures with Cusum



Number of groups = 123
Center = 80.99187
StdDev = 3.371701

Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 133

Change Detection in Summer Temperatures with Cusum

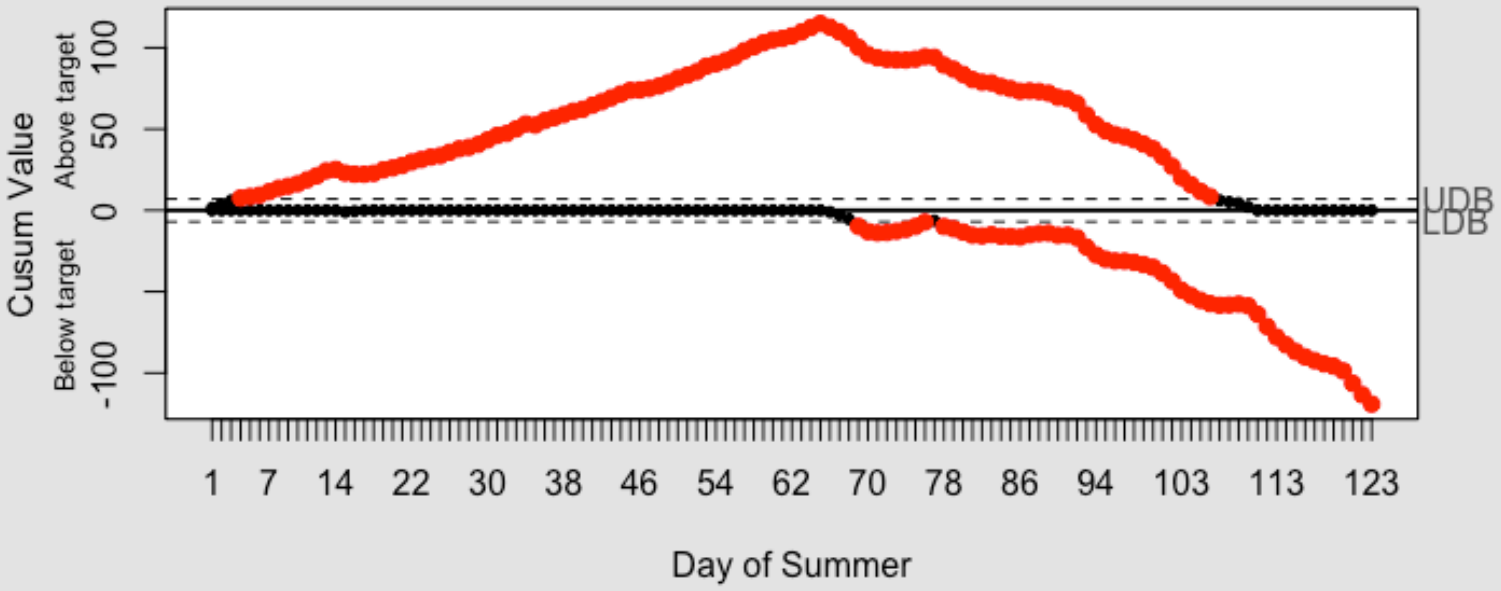


Number of groups = 123

Decision interval (std. err.) = 7

StdDev = 2.601442
No. of points beyond boundaries = 132

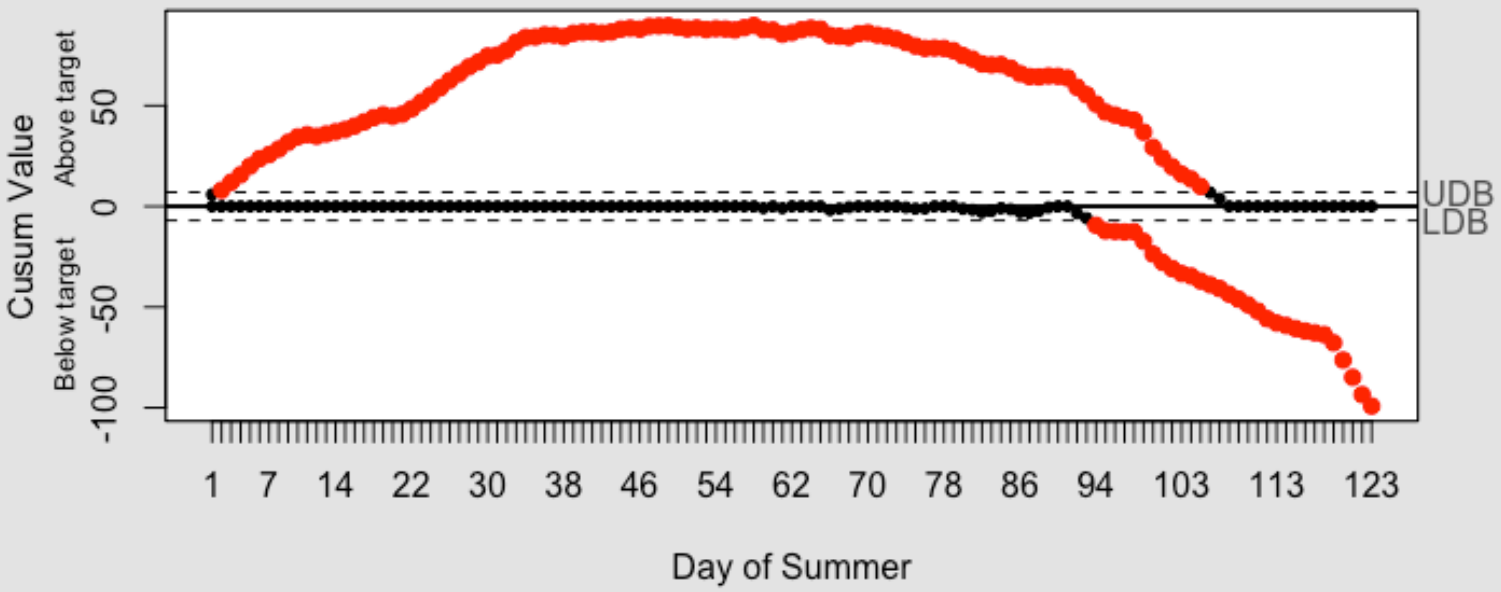
Change Detection in Summer Temperatures with Cusum



Number of groups = 123
Center = 85.27642
StdDev = 3.190036

Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 157

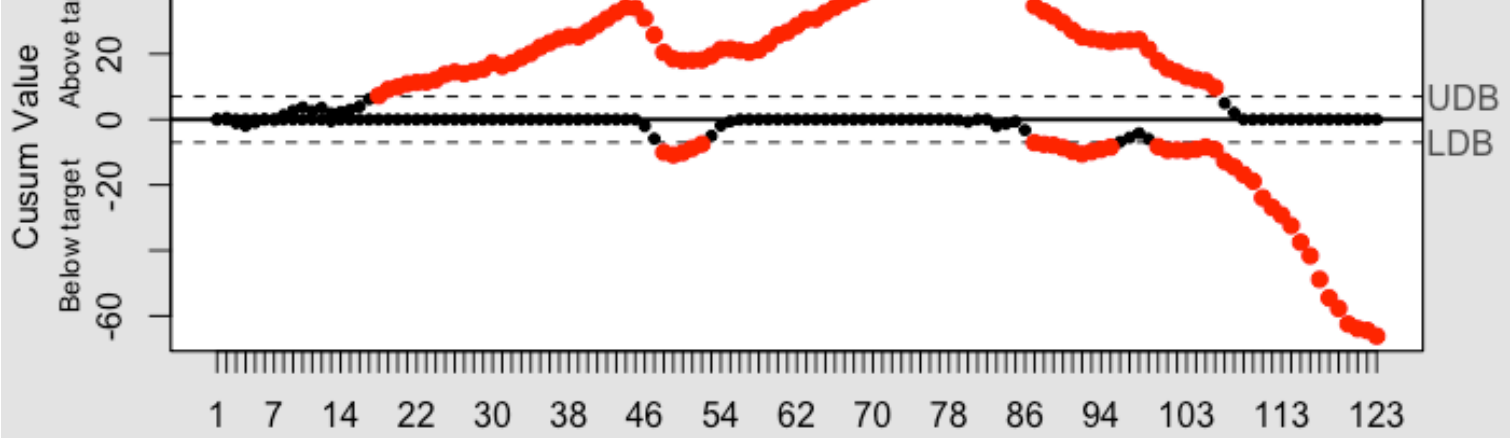
Change Detection in Summer Temperatures with Cusum



Number of groups = 123
Center = 84.65041
StdDev = 3.124637

Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 134

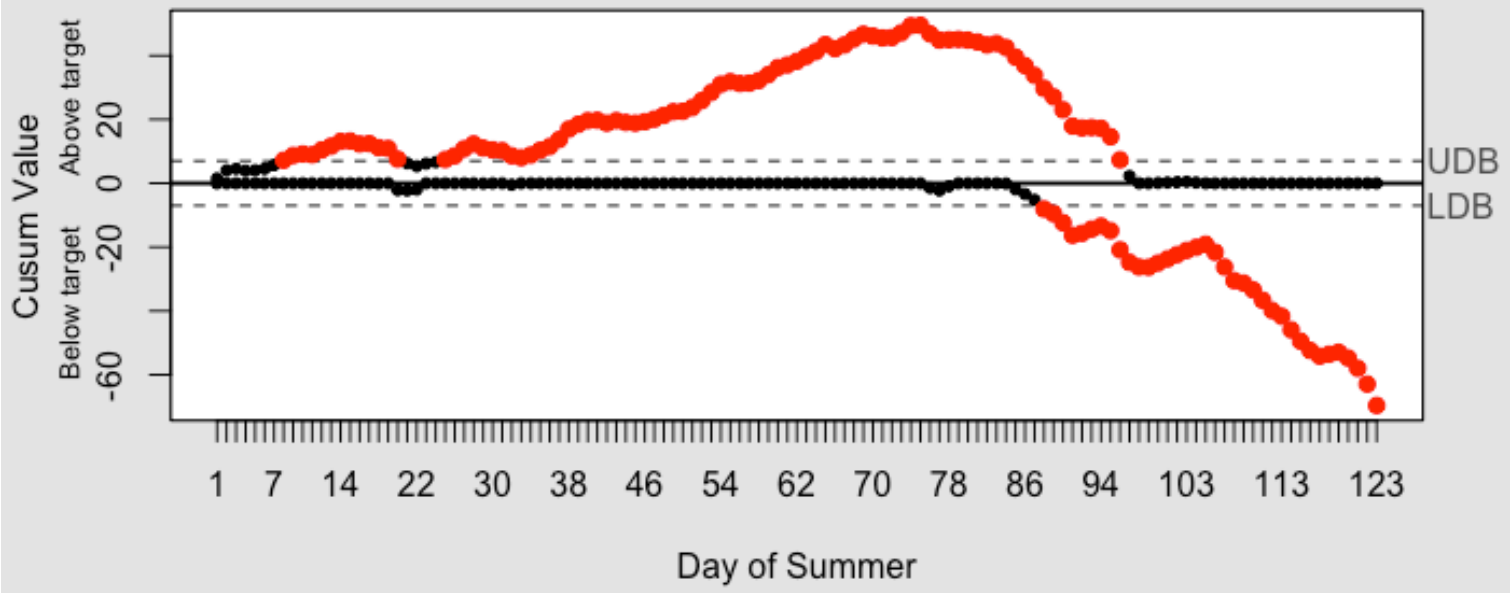
Change Detection in Summer Temperatures with Cusum



Number of groups = 123
Center = 81.66667
StdDev = 3.299035

Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 127

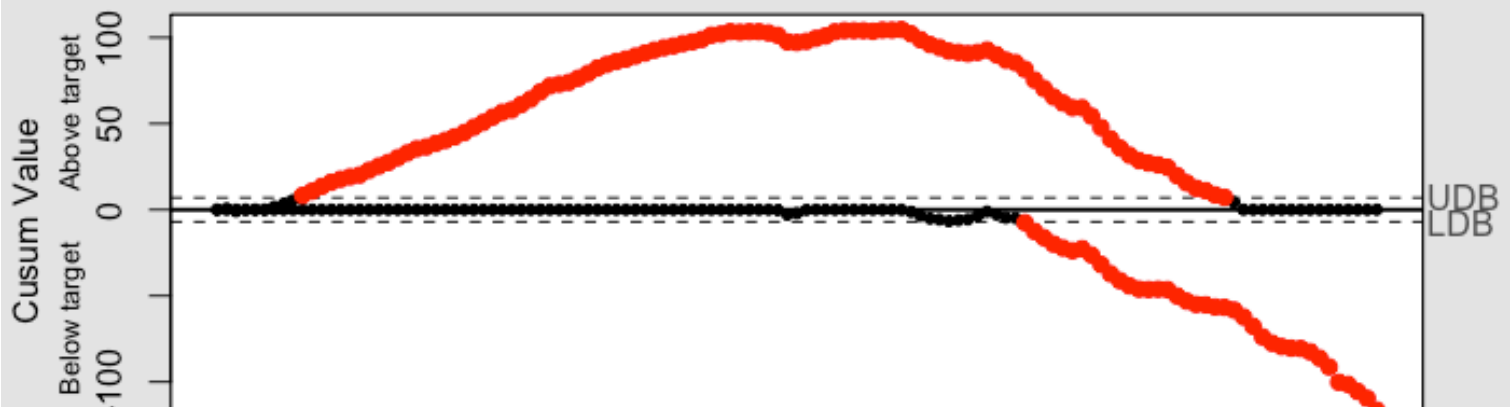
Change Detection in Summer Temperatures with Cusum

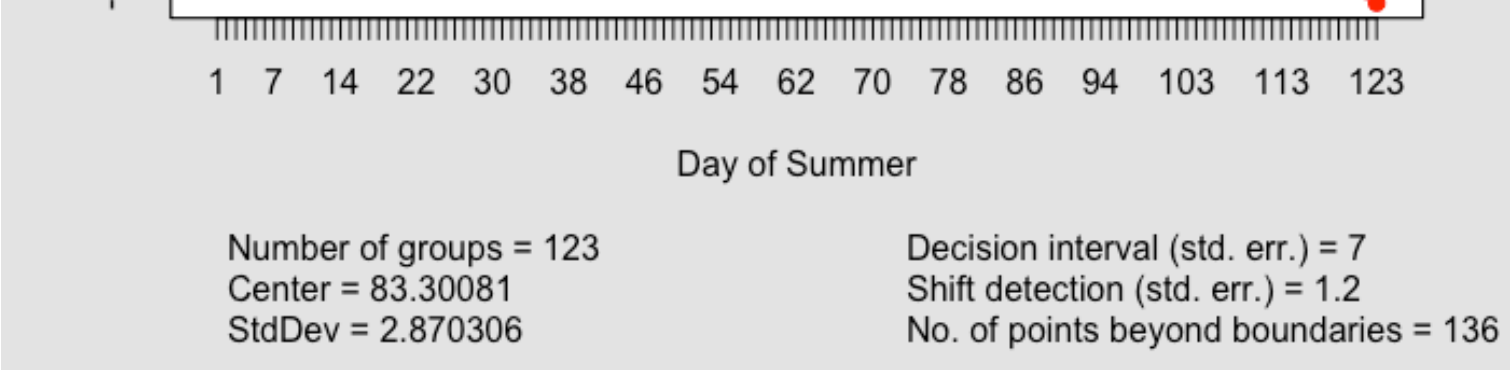


Number of groups = 123
Center = 83.94309
StdDev = 2.870306

Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 121

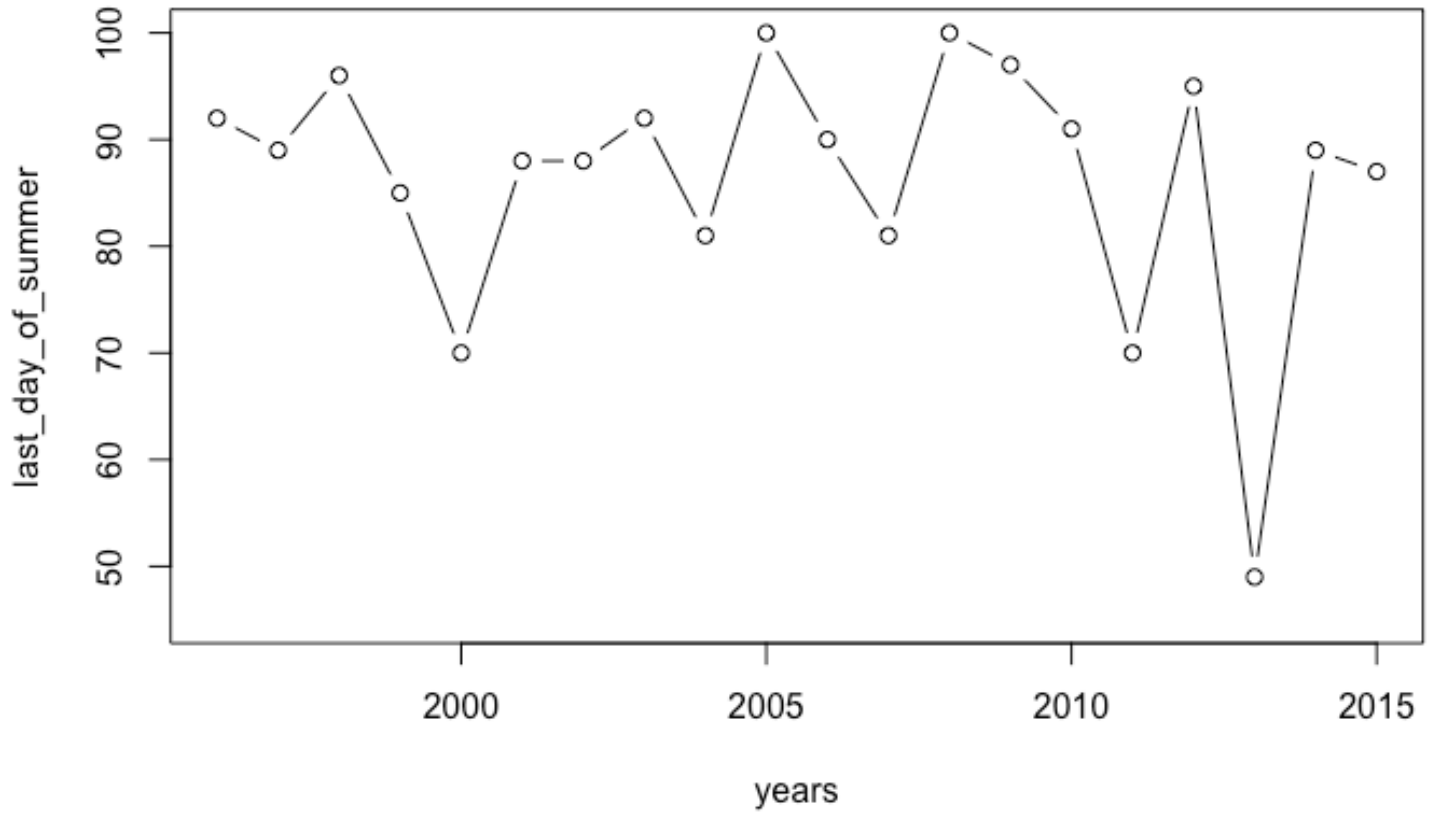
Change Detection in Summer Temperatures with Cusum





Hide

```
years <- seq(from = 1996, to =2015)
plot(years, last_day_of_summer, type = "b", ylim = c(45,100))
```



To test these results I repeated the cusum function but this time charted the end of summer as the day in which temperatures decreased from then on (i.e. from that point forward each daily temperature was below the mu). I did this with the mu being that summer's average temperature and also the population mu (avg temperature across all summers).

The results were a bit different in that the cusum version above generally marked the end of summer as later than the true last day of summer. This could be a result of the stringent threshold. The above version also thought the end of the 2013 summer to be very early (~day 50) which was the result of a brief period of unusually cool weather. Overall I think the cusum version did well.

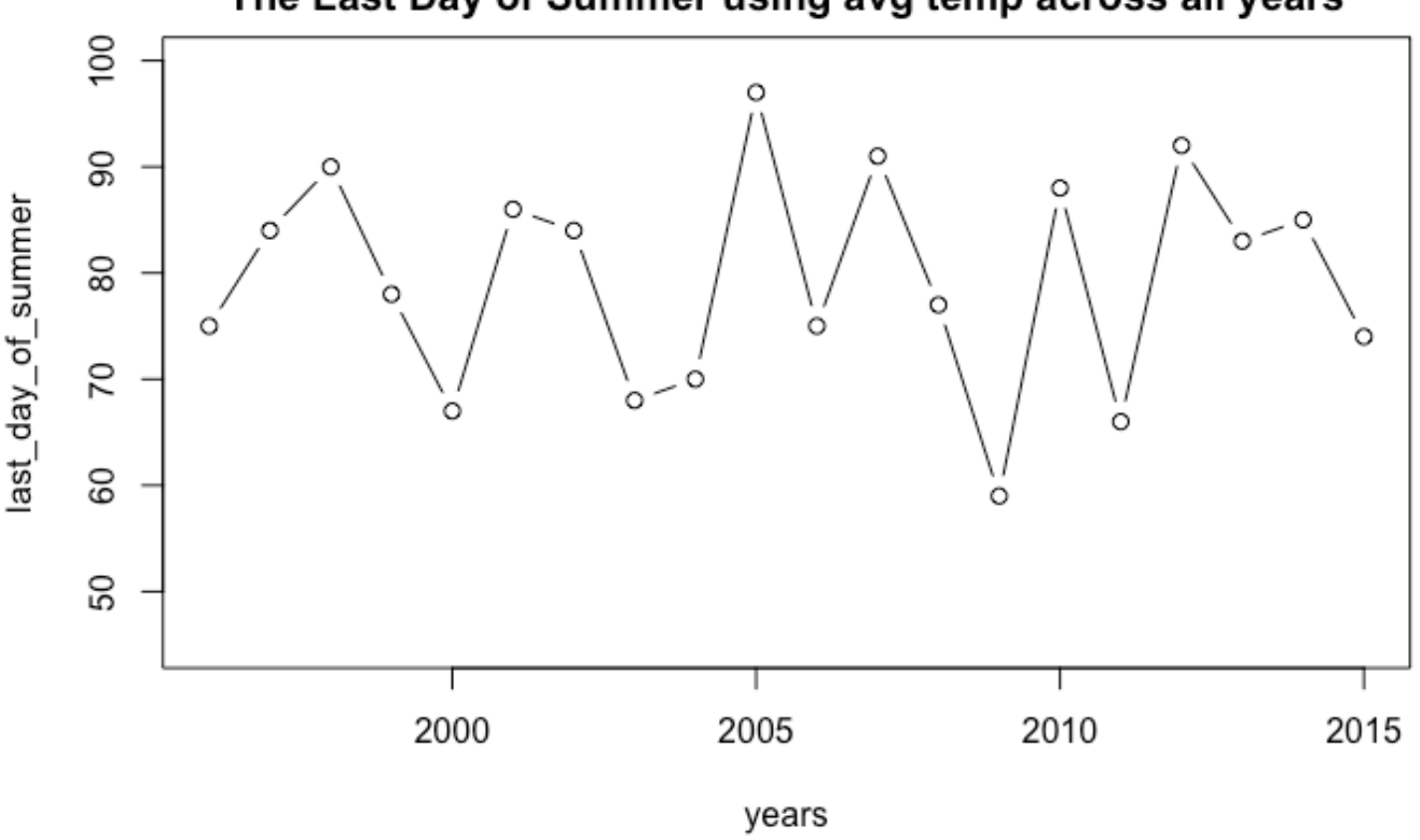
Hide

```
last_day_of_summer <- c()
for (year in 2:21){
  temp_changes <- with(cusum(temps[,year], center = 83.33902, plot =
FALSE),
                       cbind(data, "Ci+" = pos, "neg" = -neg))
  # Day of summer
  day_count <-  1

  # vector of days where temperature is NOT increasing
  temp_increase_day <- c()

  # Loop to find the last day in which the temperature increases (i.e
. daily temp - mu is negative)

  # Looping through the cusum data
  for (summer_day in seq(nrow(temp_changes))){

    # If the temperature decrease is 0 (i.e. temperature is still inc
reasing)
    # append that day number to our temp_increase_day variable
    if (temp_changes[summer_day,"neg"]==0){
      temp_increase_day <- append(temp_increase_day, day_count)
    }
    # continute on to the next day
    day_count <- day_count + 1
  }

  # The day after the last day in which temperature is increasing
  # Signifies the end of summer (temperatures are now decreasing from
this point)
  last_day_of_summer <- append(last_day_of_summer, max(temp_increase_
day)+1)

}
last_day_of_summer
```

```
 [1] 75 84 90 78 67 86 84 68 70 97 75 91 77 59 88 66 92 83 85 74
```

Hide

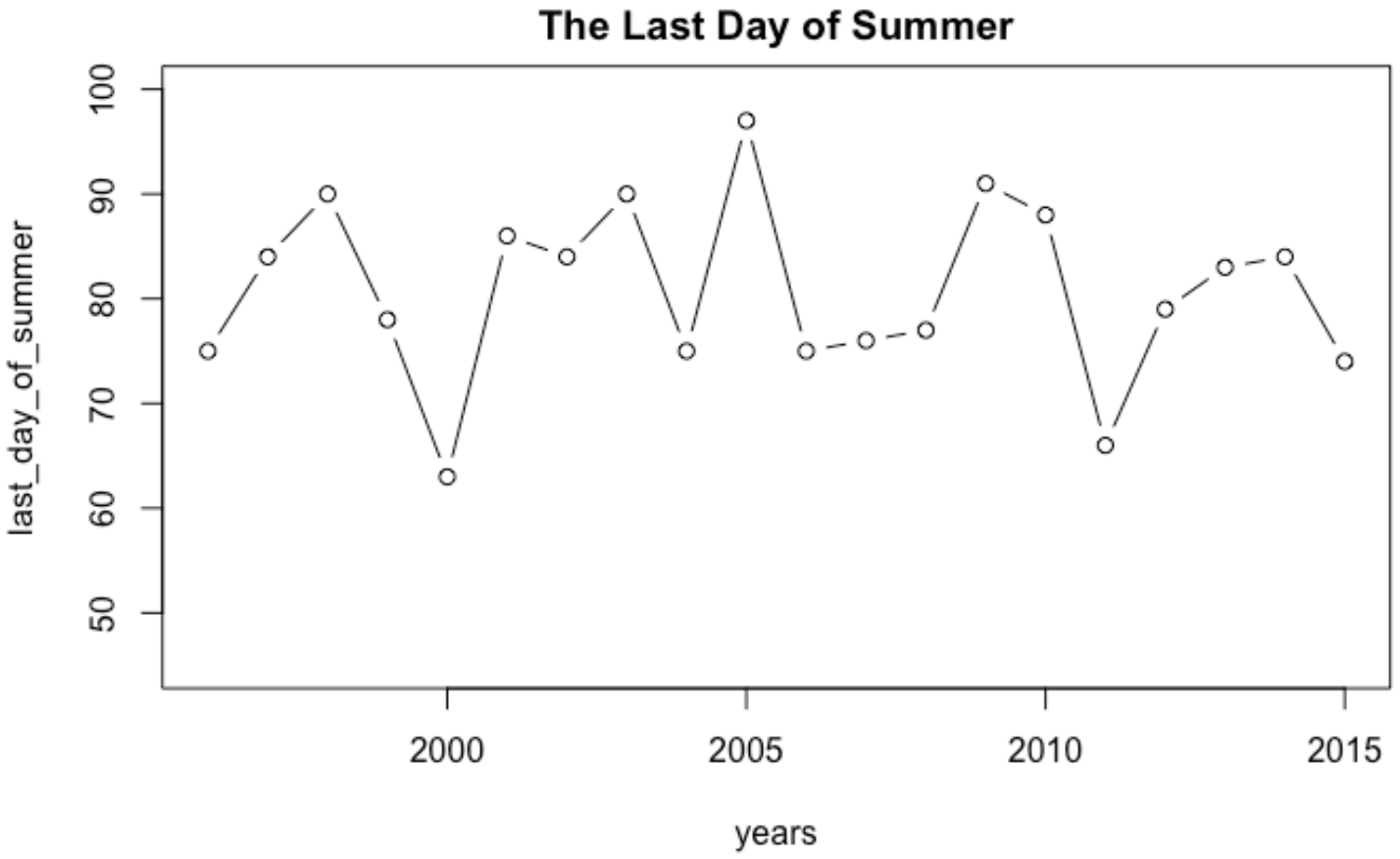
```
years <- seq(from = 1996, to =2015)
plot(years, last_day_of_summer, type = "b", ylim = c(45,100), main =
"The Last Day of Summer using avg temp across all years")
```



Hide

```
last_day_of_summer <- c()
for (year in 2:21){
  temp_changes <- with(cusum(temps[,year], plot = FALSE),
                       cbind(data, "Ci+" = pos, "neg" = -neg))
  # Day of summer
  day_count <-  1

  # vector of days where temperature is NOT increasing
  temp_increase_day <- c()

  # Loop to find the last day in which the temperature increases (i.e
. daily temp - mu is negative)

  # Looping through the cusum data
  for (summer_day in seq(nrow(temp_changes))){

    # If the temperature decrease is 0 (i.e. temperature is still inc
reasing)
    # append that day number to our temp_increase_day variable
    if (temp_changes[summer_day,"neg"]==0){
      temp_increase_day <- append(temp_increase_day, day_count)
    }
    # continute on to the next day
    day_count <- day_count + 1
  }
```

```
  # The day after the last day in which temperature is increasing
  # Signifies the end of summer (temperatures are now decreasing from
this point)
  last_day_of_summer <- append(last_day_of_summer, max(temp_increase_
day)+1)

}
years <- seq(from = 1996, to =2015)
plot(years, last_day_of_summer, type = "b", ylim = c(45,100), main =
"The Last Day of Summer")
```

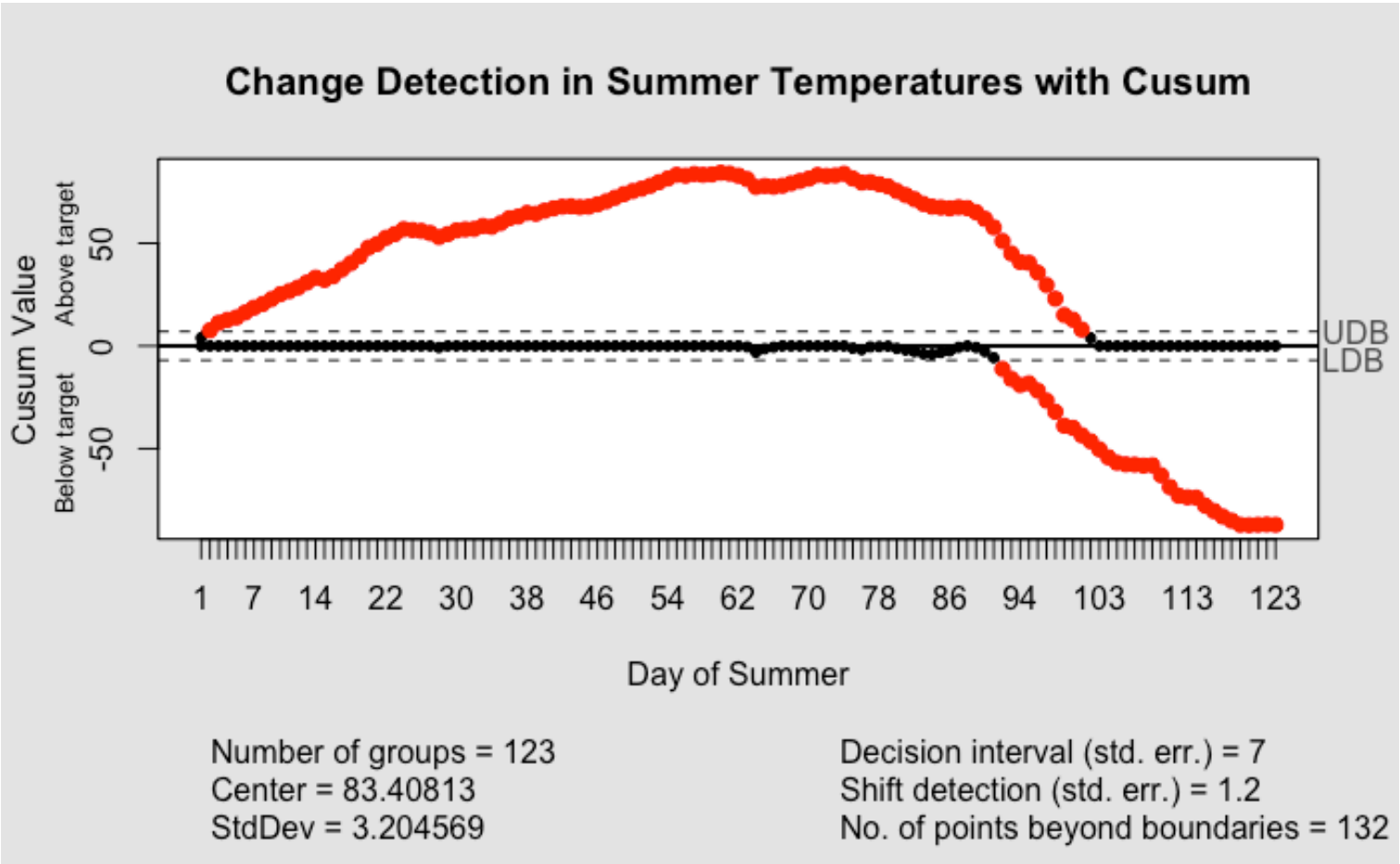


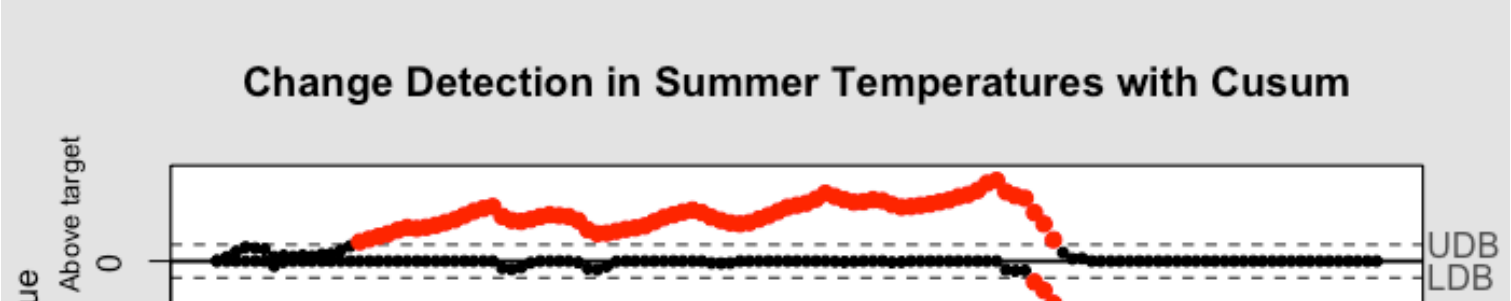
Question #3.1: CUSUM for Summer's End To investigate whether Atlanta's summer climate increased over the years using the cusum method I chose to use the average summer temperature of the first 5 years as mu. If the climate was getting warmer, cusum would detect a postive change earlier and longer.

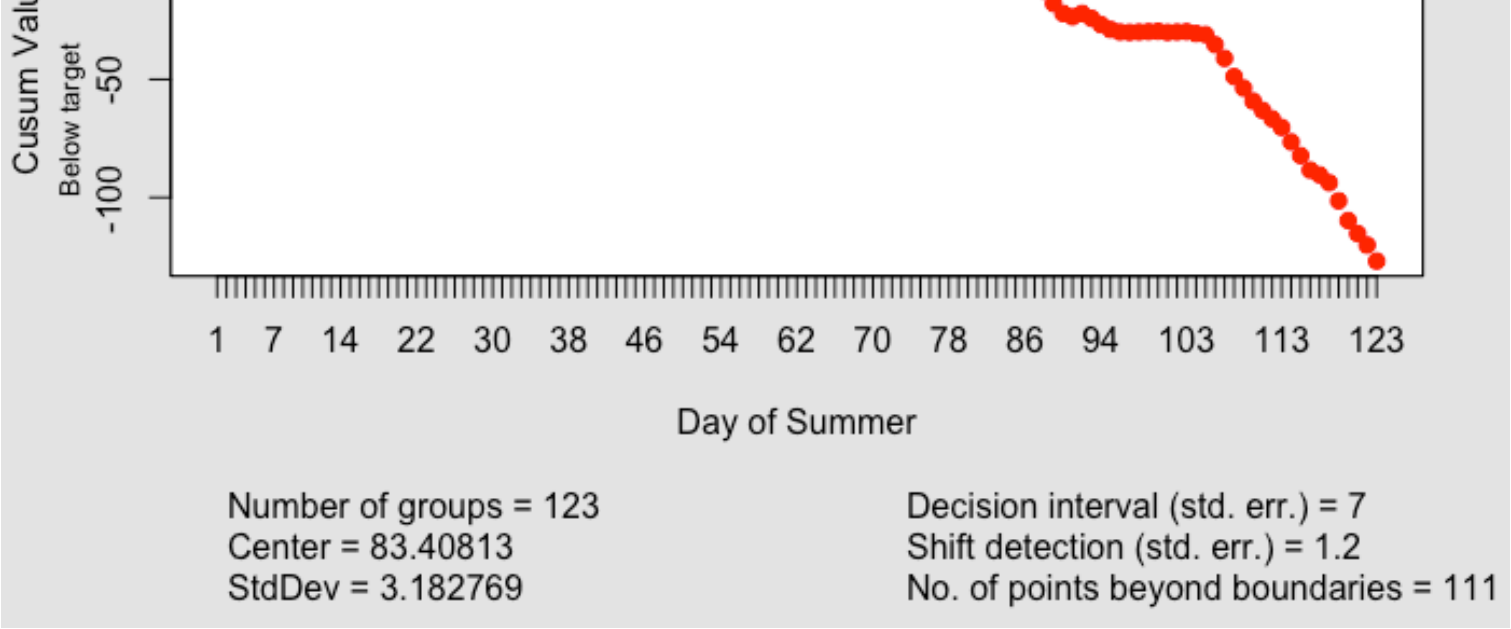
Hide

```
# First five years of temperatures
five_years <- temps[,2:6]
#
five_years <- transform(five_years, sum=rowMeans(five_years))
five_year_summer_avg <- mean(five_years[,6])
for (year in 2:21){
  # Using each year's temps for cusum
  years_cusum <- cusum(temps[,year],
  # The center is the average of the temperatures. The below value is
```

```
  # The center is the average of the temperatures. The below value is
the average across
  # all years. If it isn't used, the cusum function takes that summer
's average temperature
  center = five_year_summer_avg,

  # Decision interval is the threshold for when a change is detected

  # Will use 7
  decision.interval = 7,

  # The critical value (our senstivity to change). Will use 1.2
  se.shift = 1.2,

  # Chart name
  data.name = "Summer Temperatures",
  title = "Change Detection in Summer Temperatures with Cusum",
  xlab = "Day of Summer",
  ylab = "Cusum Value")
}
```



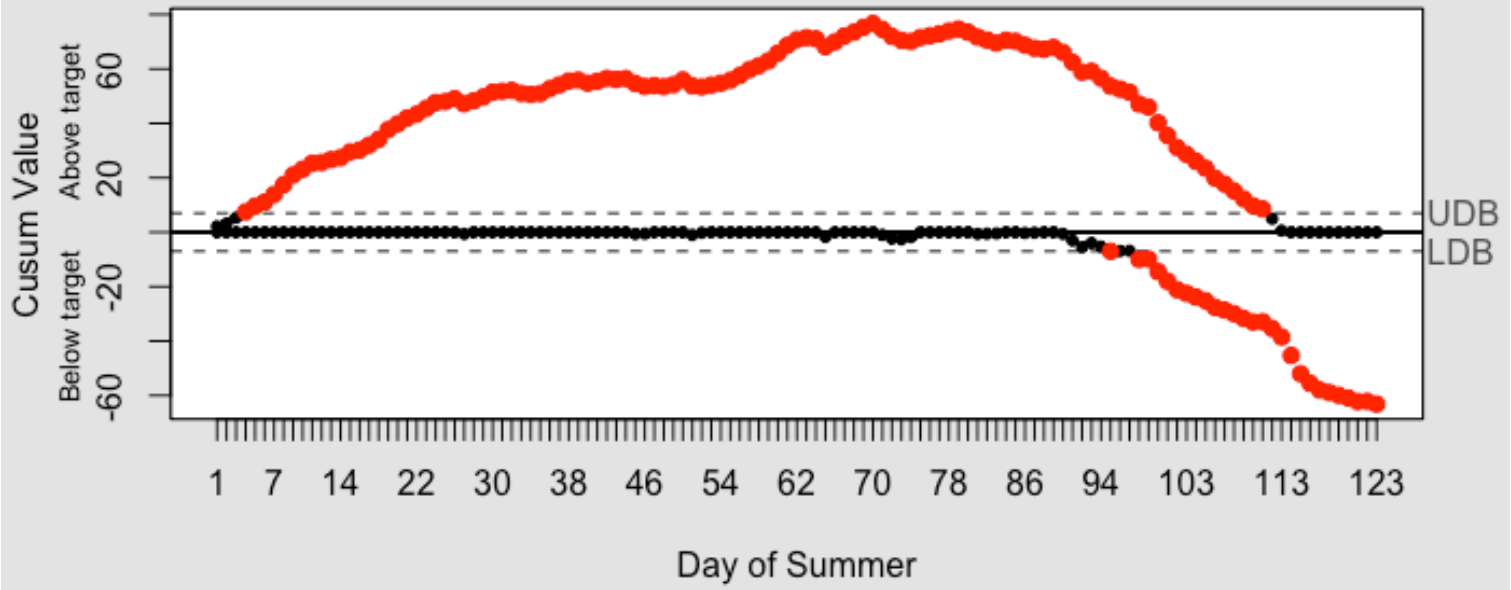




Number of groups = 123
Center = 83.40813
StdDev = 3.182769

Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 111

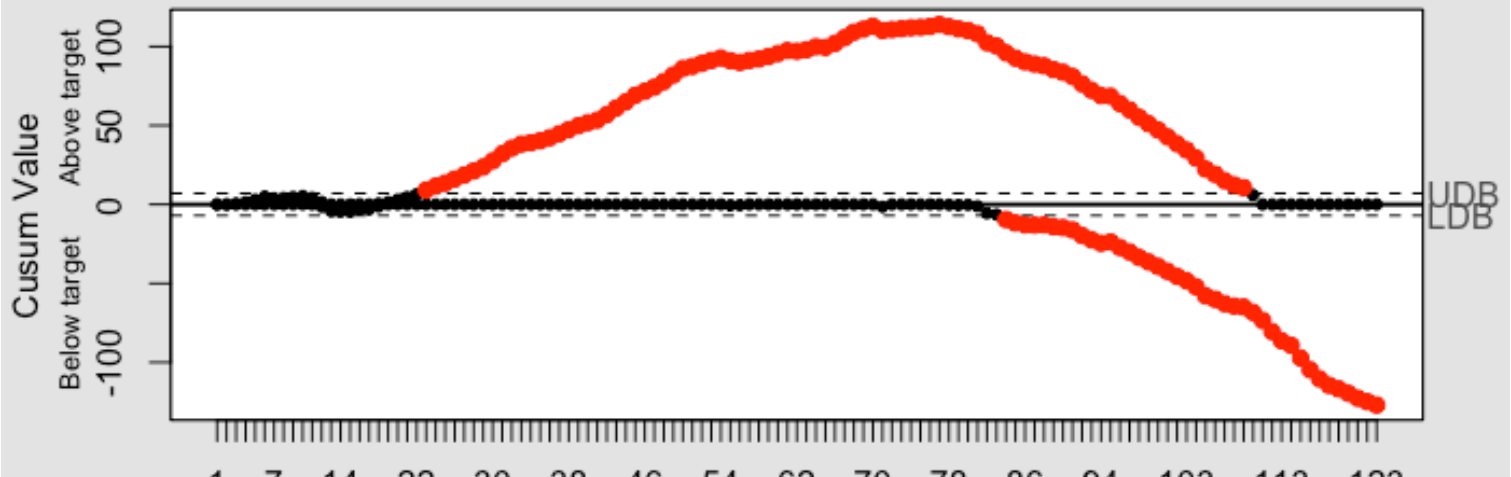
Change Detection in Summer Temperatures with Cusum



Number of groups = 123
Center = 83.40813
StdDev = 2.77584

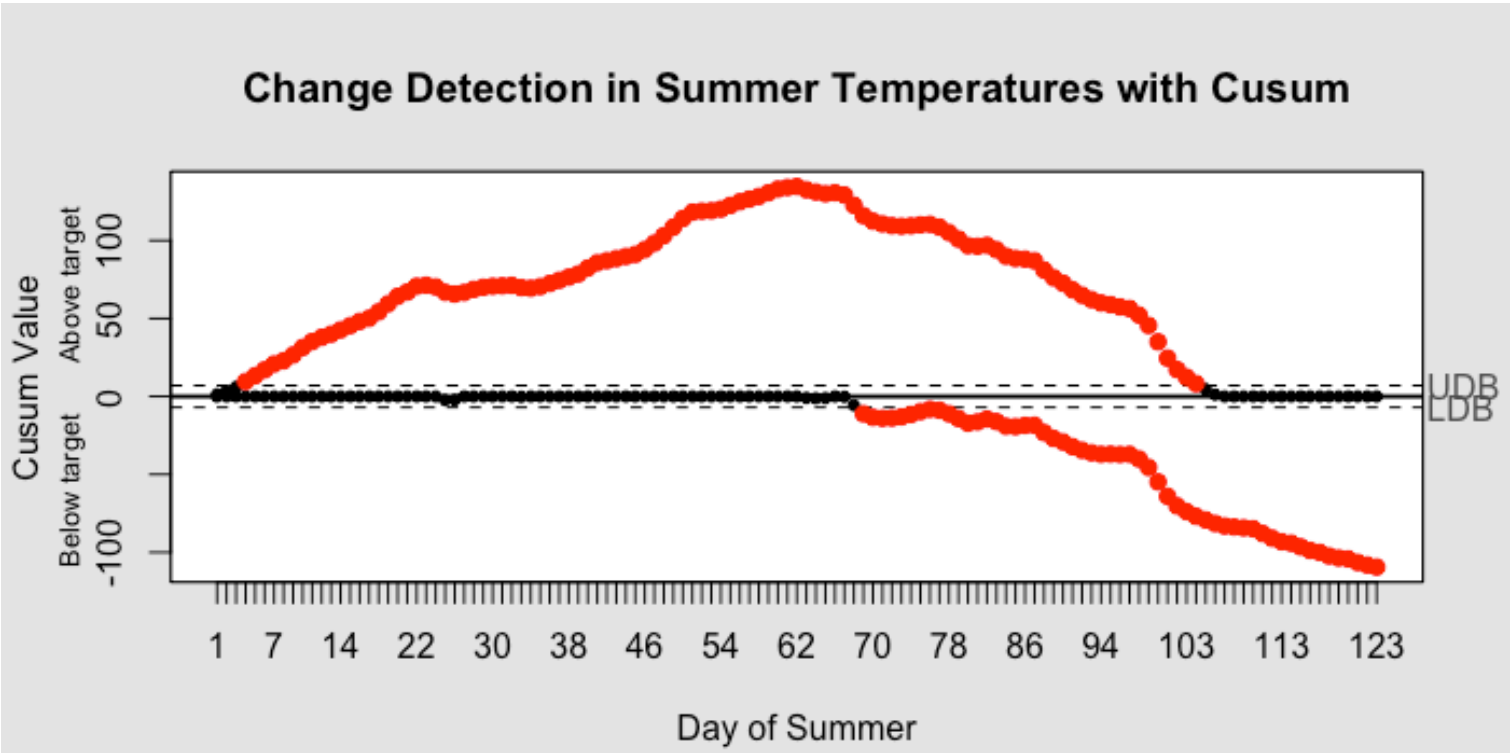
Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 135

Change Detection in Summer Temperatures with Cusum



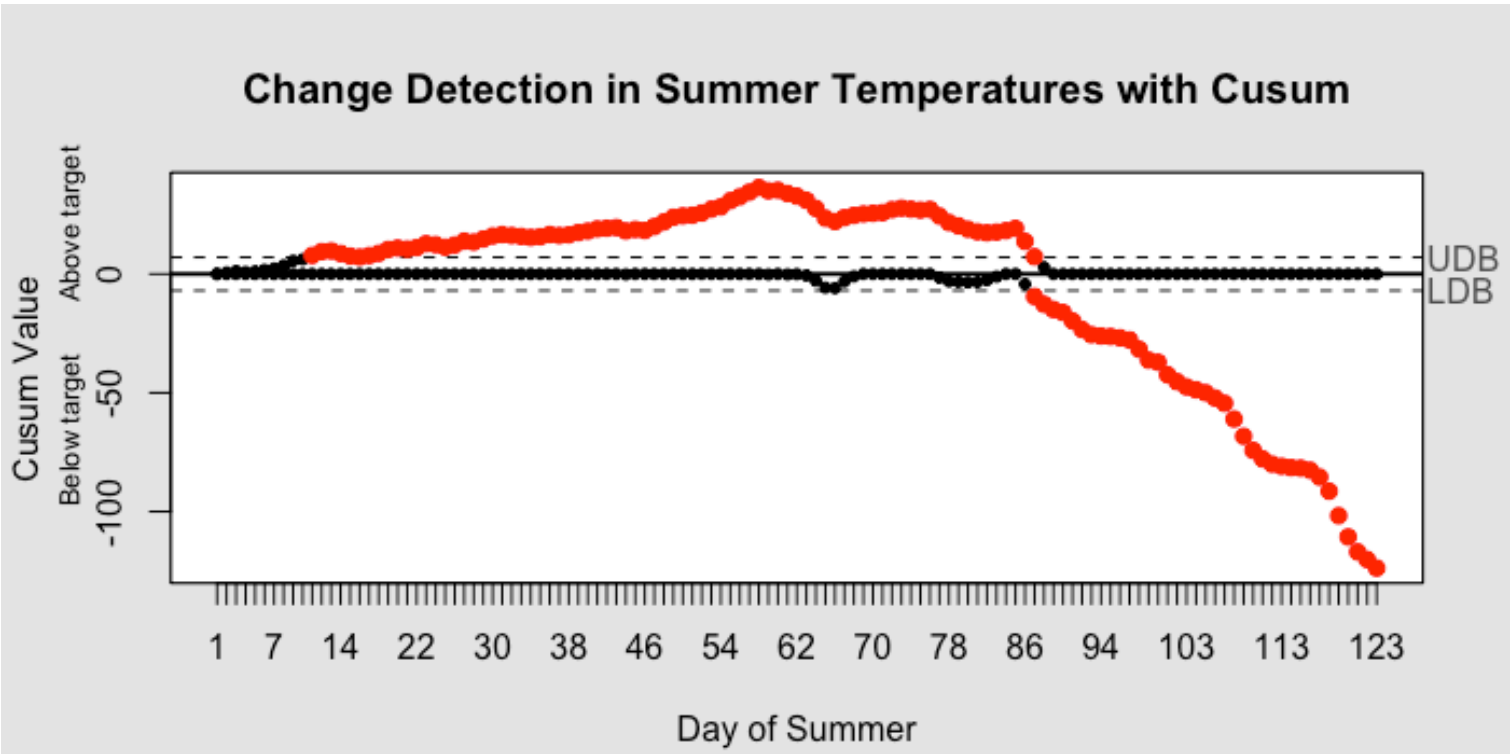
Number of groups = 123
Center = 83.40813
StdDev = 3.044704

Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 127



Number of groups = 123
Center = 83.40813
StdDev = 2.884839

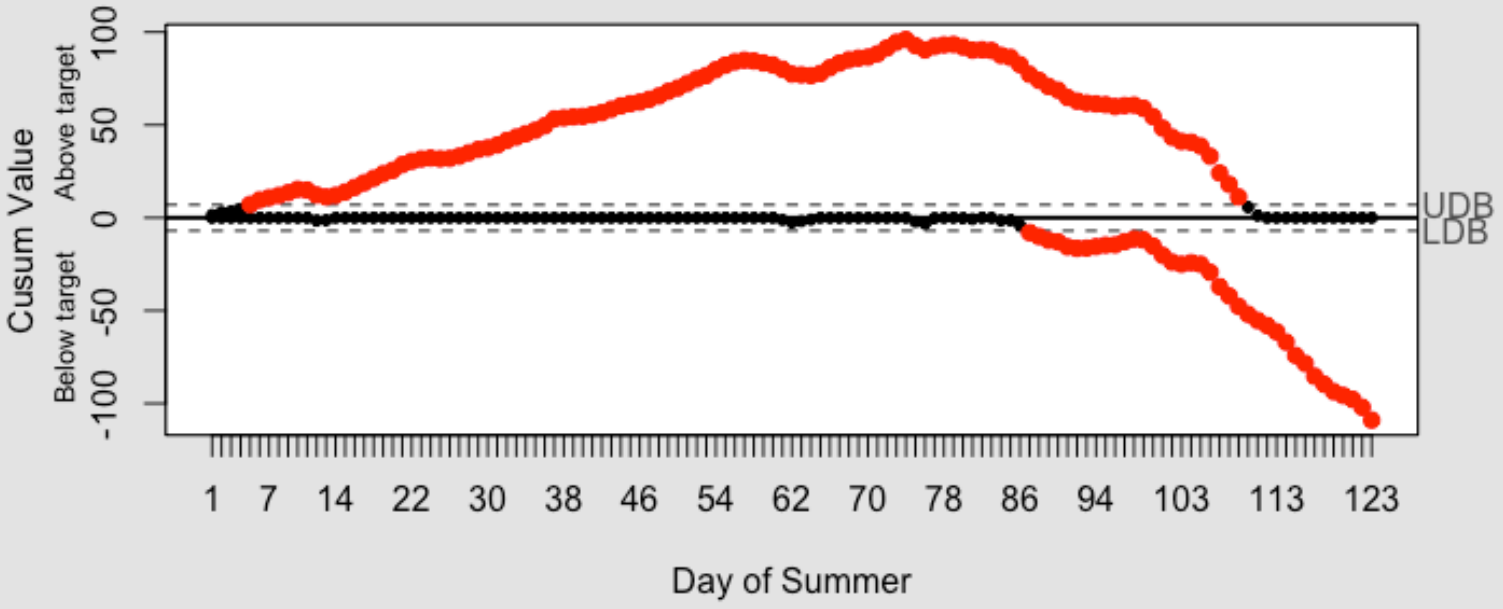
Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 156



Number of groups = 123
Center = 83.40813
StdDev = 2.972038

Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 114

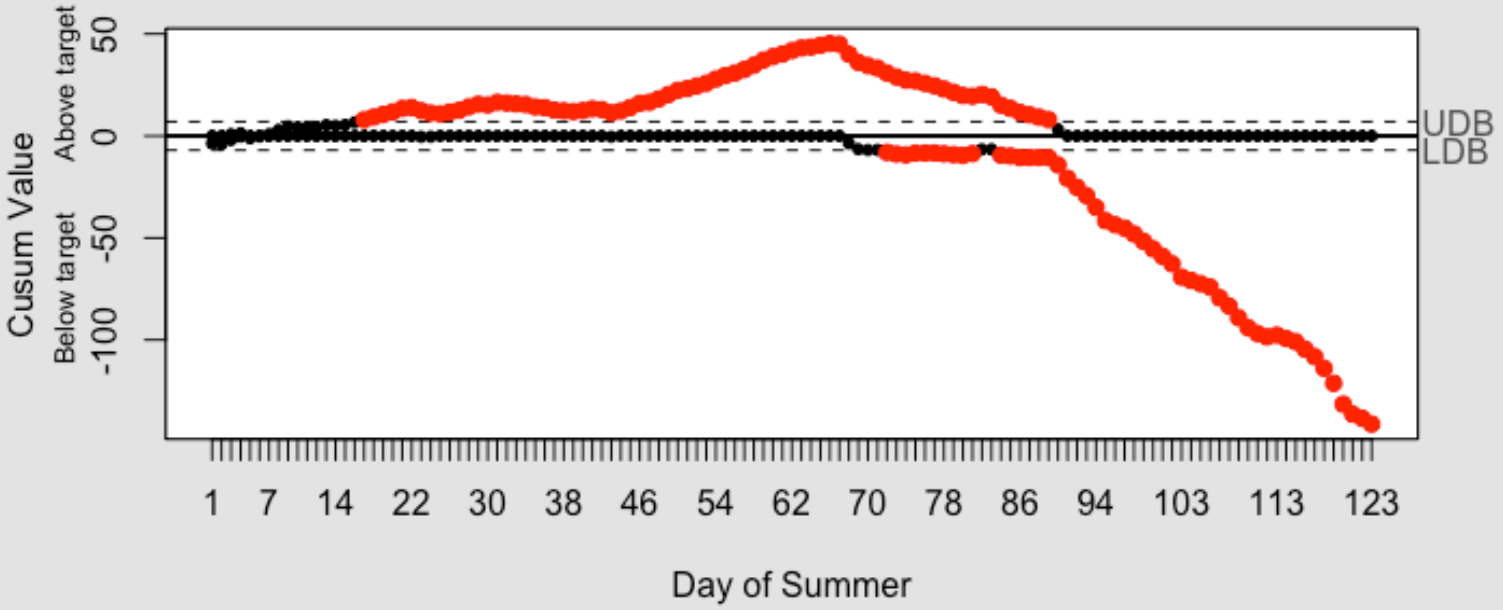
Change Detection in Summer Temperatures with Cusum



Number of groups = 123
Center = 83.40813
StdDev = 3.13917

Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 142

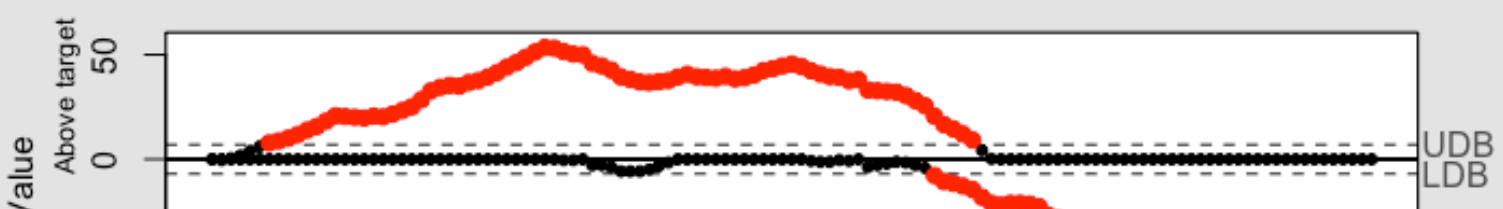
Change Detection in Summer Temperatures with Cusum

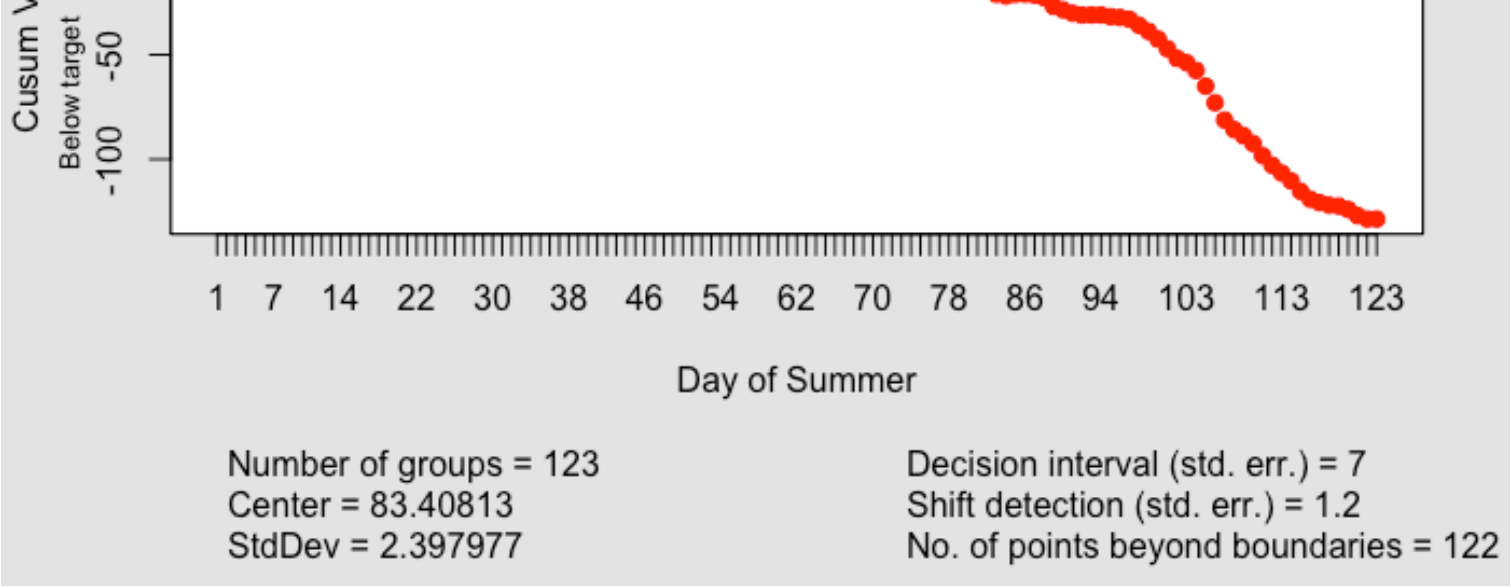


Number of groups = 123
Center = 83.40813
StdDev = 2.441577

Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 123

Change Detection in Summer Temperatures with Cusum

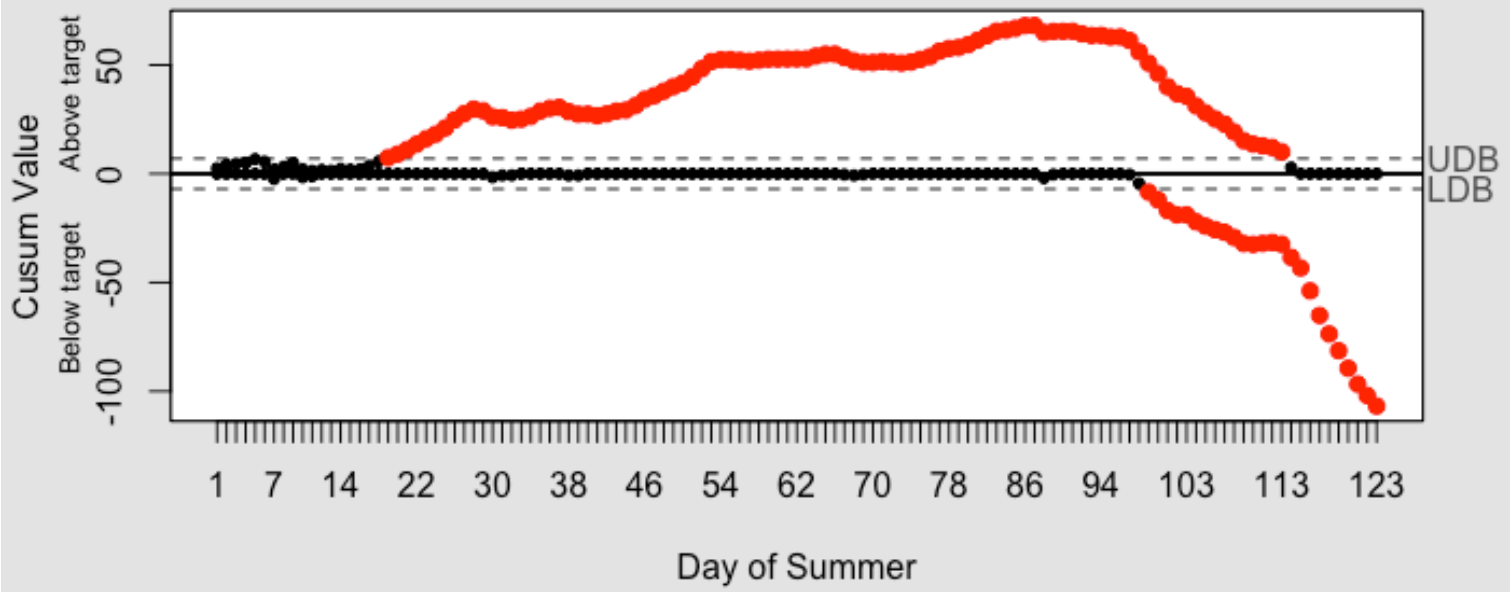




Number of groups = 123
Center = 83.40813
StdDev = 2.397977

Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 122

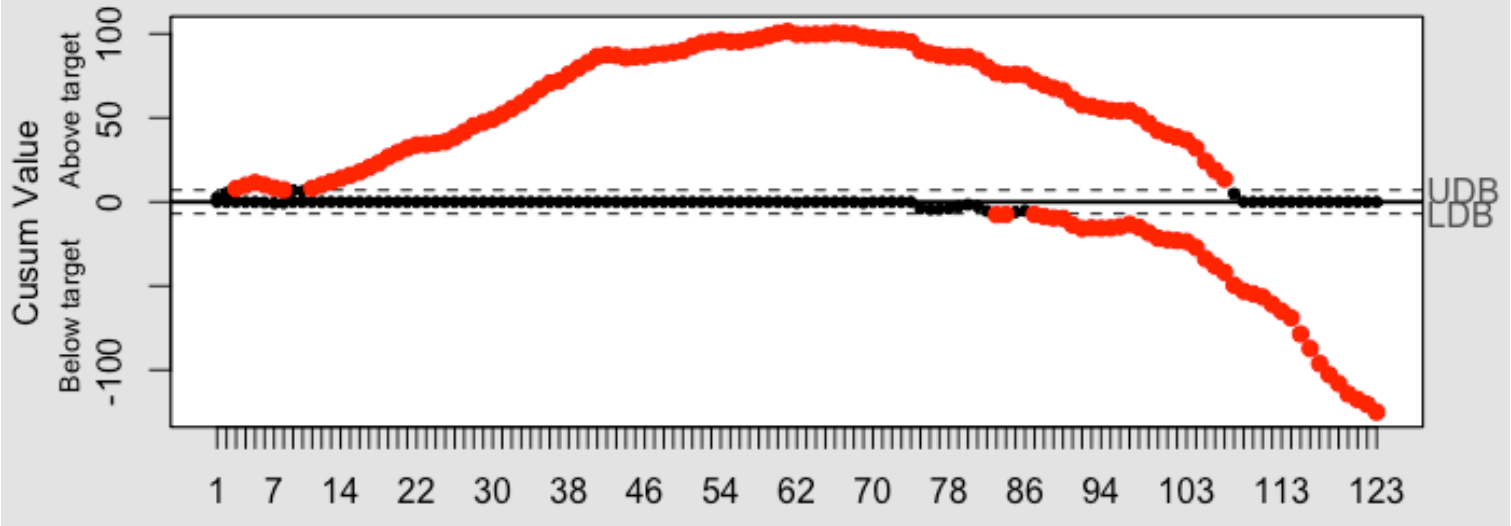
Change Detection in Summer Temperatures with Cusum



Number of groups = 123
Center = 83.40813
StdDev = 2.463376

Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 120

Change Detection in Summer Temperatures with Cusum

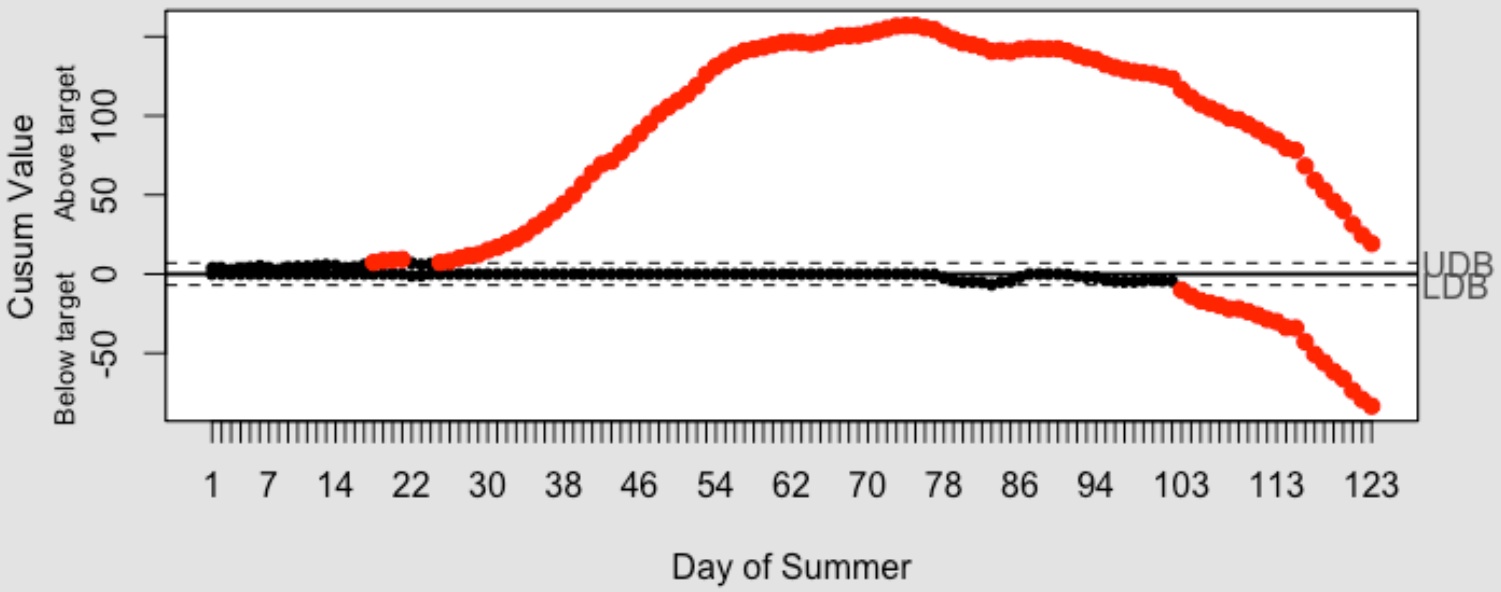


Day of Summer

Number of groups = 123
Center = 83.40813
StdDev = 2.972038

Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 142

Change Detection in Summer Temperatures with Cusum

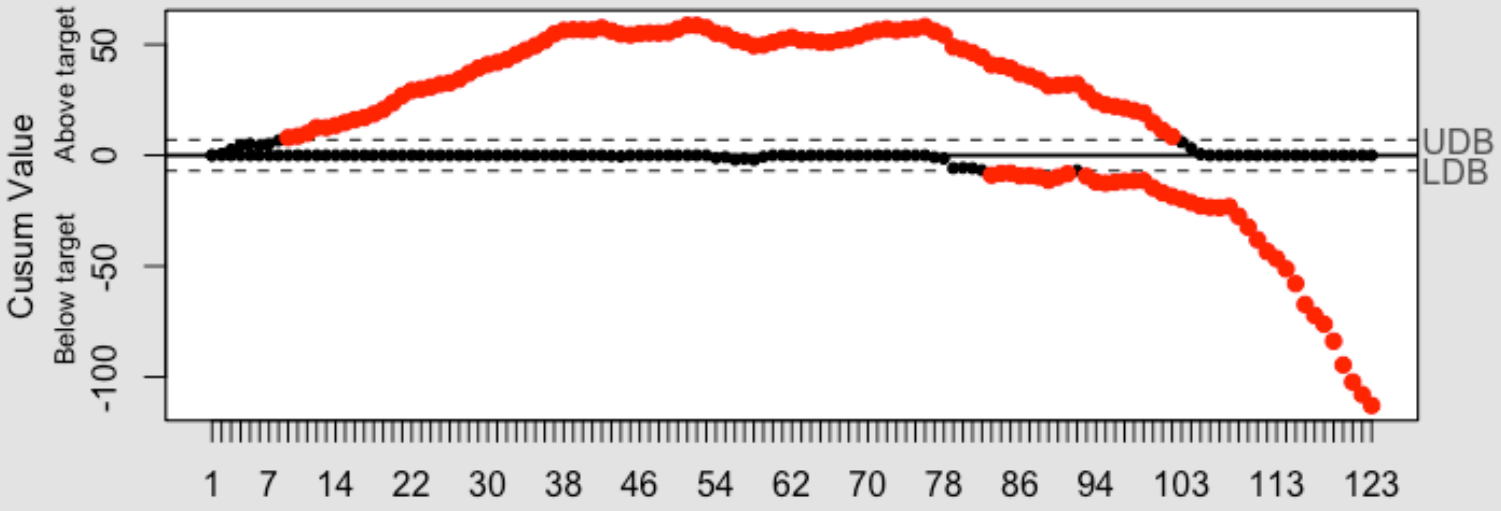


Day of Summer

Number of groups = 123
Center = 83.40813
StdDev = 2.630508

Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 124

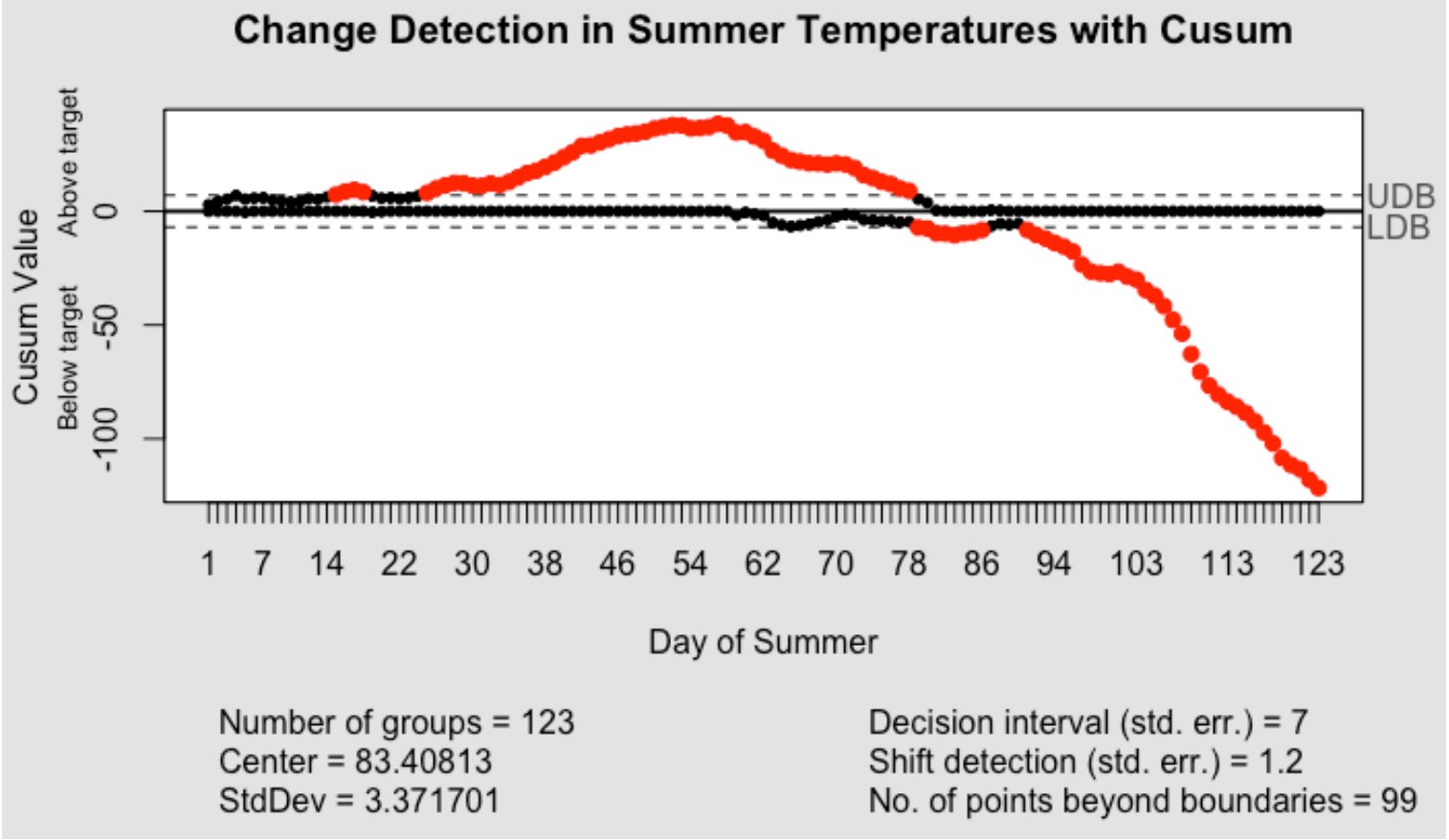
Change Detection in Summer Temperatures with Cusum



Day of Summer

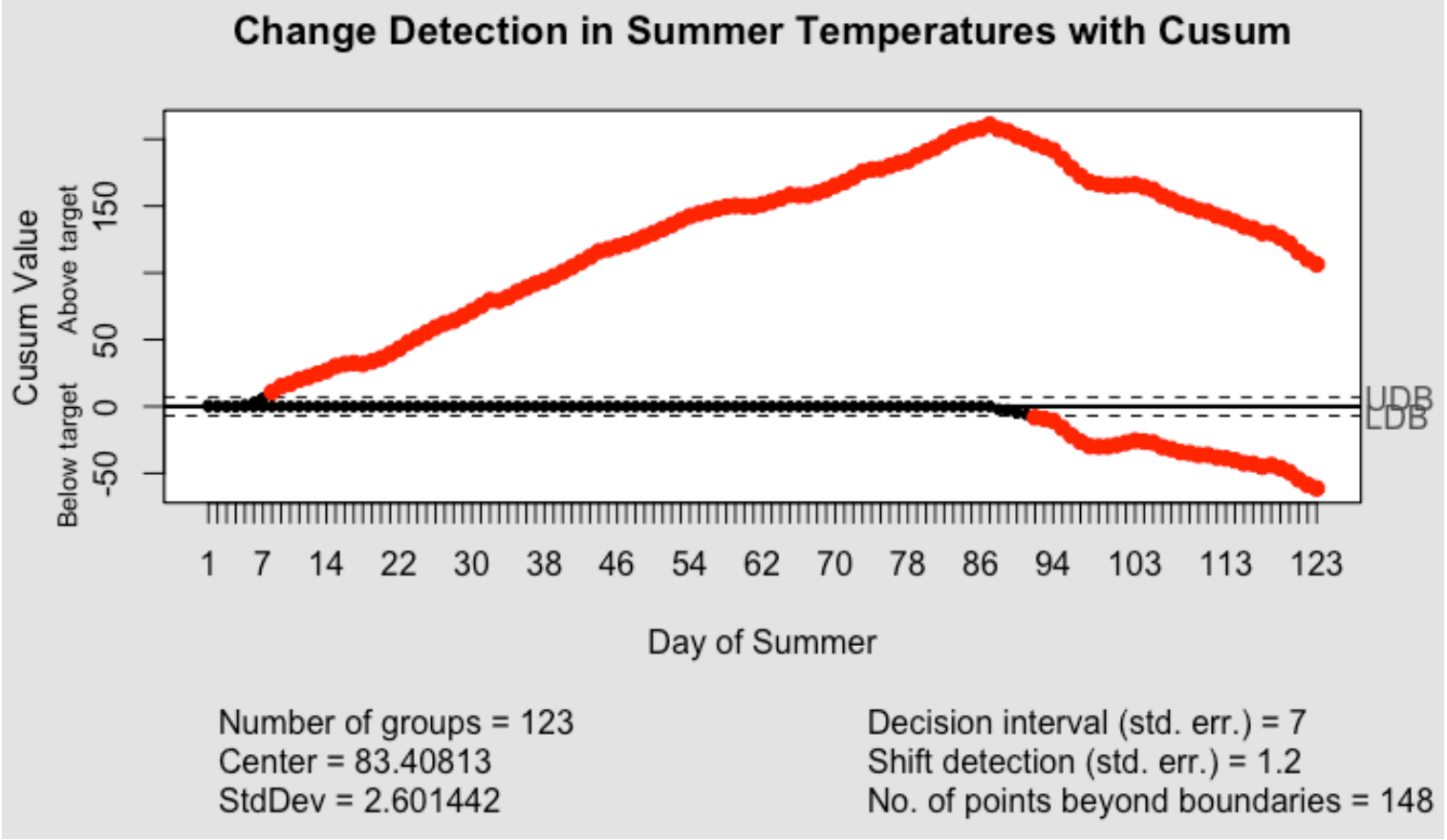
Number of groups = 123
Center = 83.40813
StdDev = 2.950238

Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 134



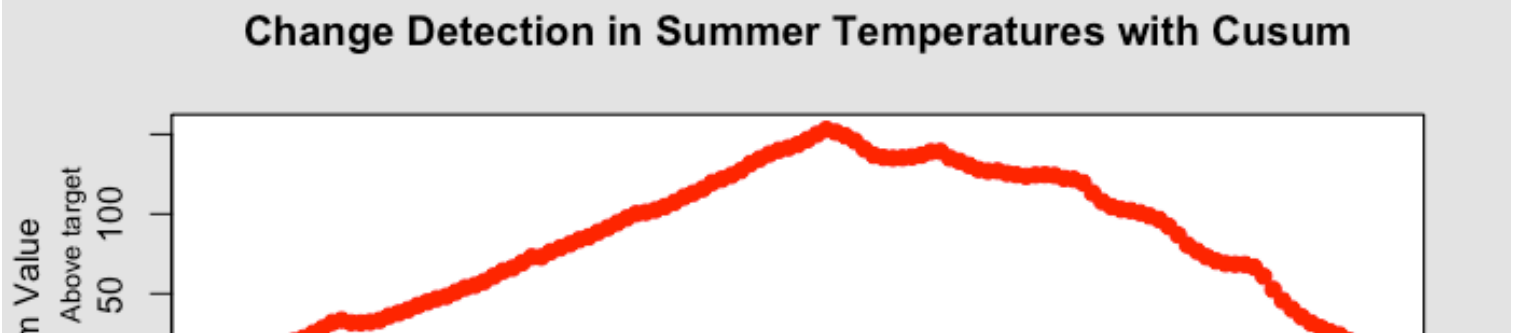
Number of groups = 123
Center = 83.40813
StdDev = 3.371701

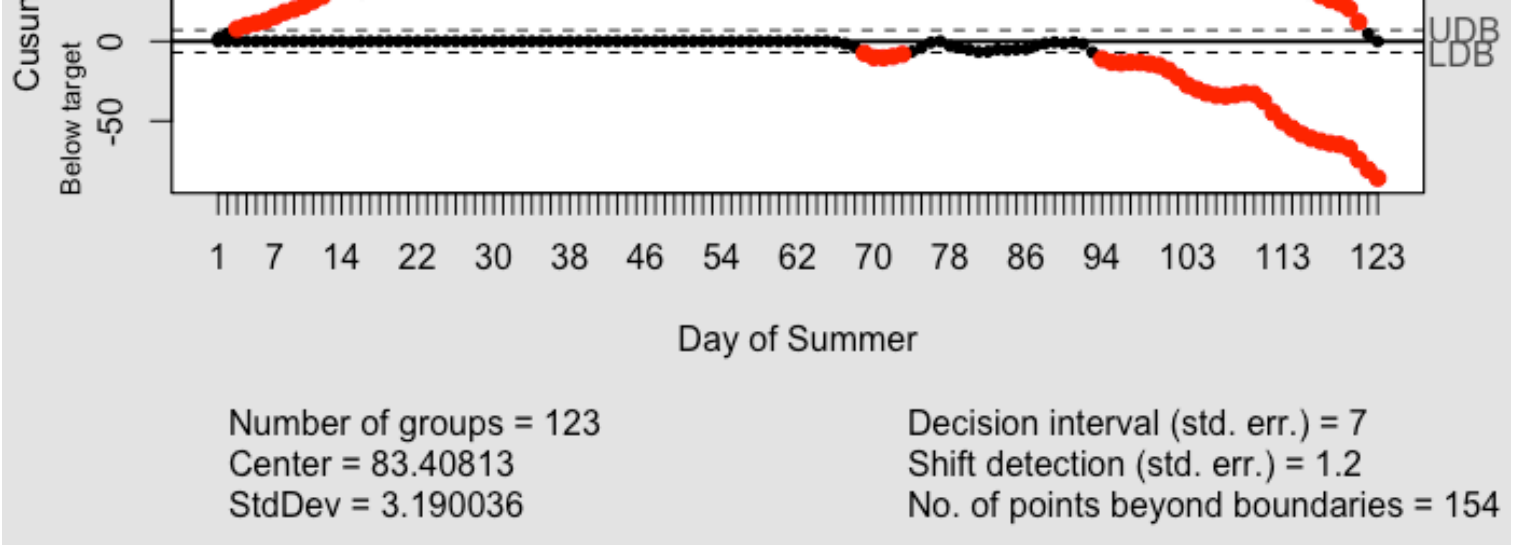
Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 99



Number of groups = 123
Center = 83.40813
StdDev = 2.601442

Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 148

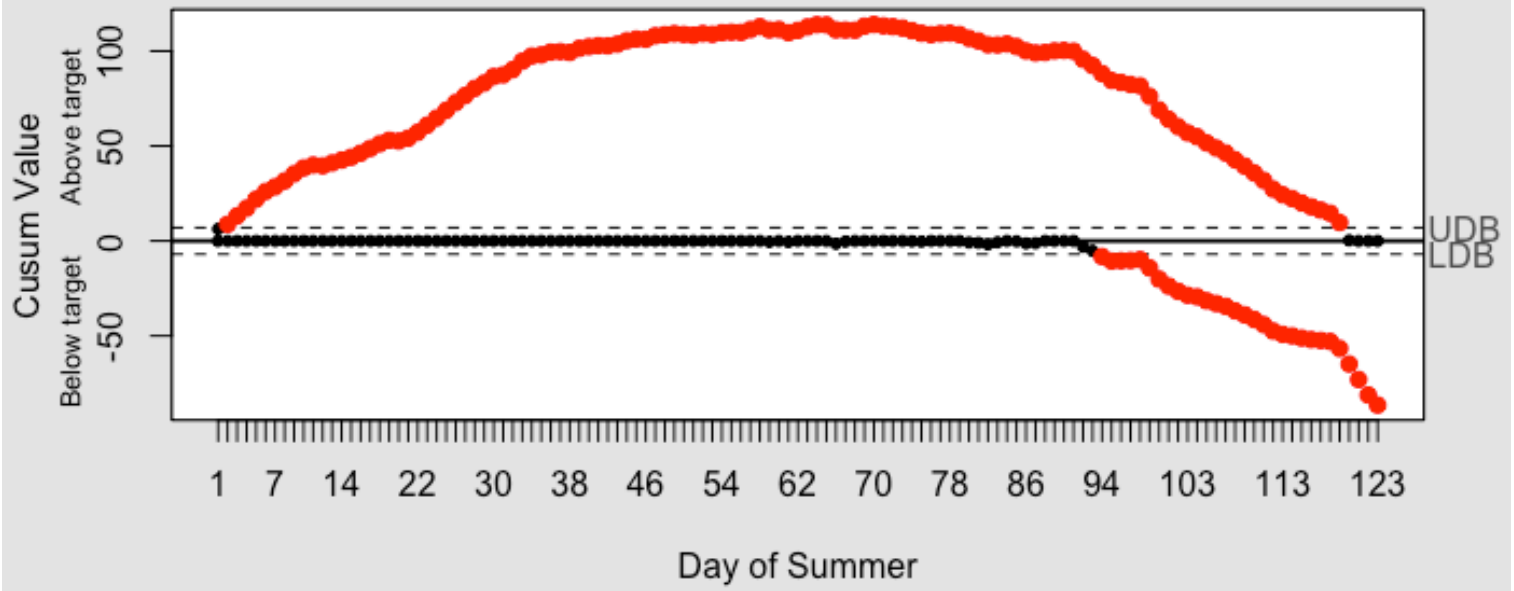




Number of groups = 123
Center = 83.40813
StdDev = 3.190036

Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 154

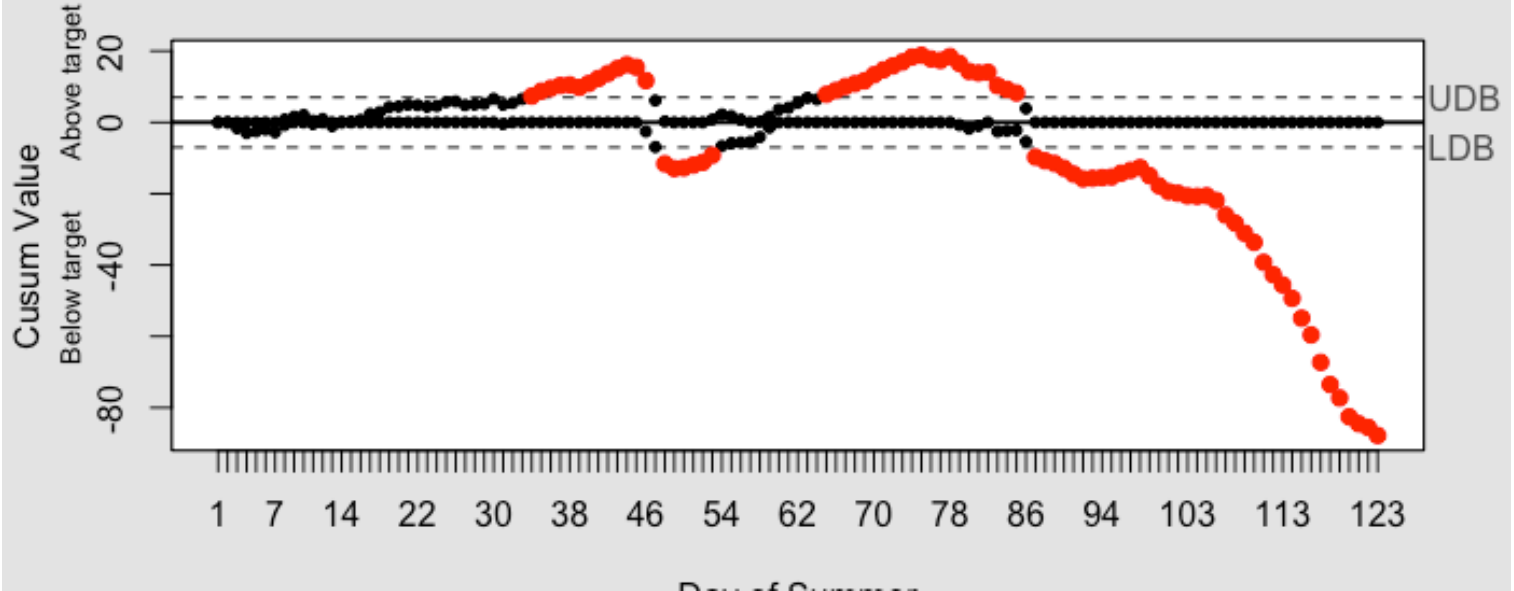
Change Detection in Summer Temperatures with Cusum

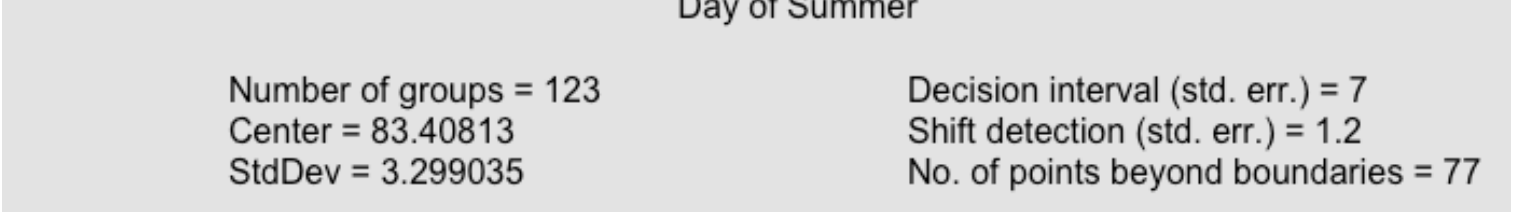


Number of groups = 123
Center = 83.40813
StdDev = 3.124637

Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 148

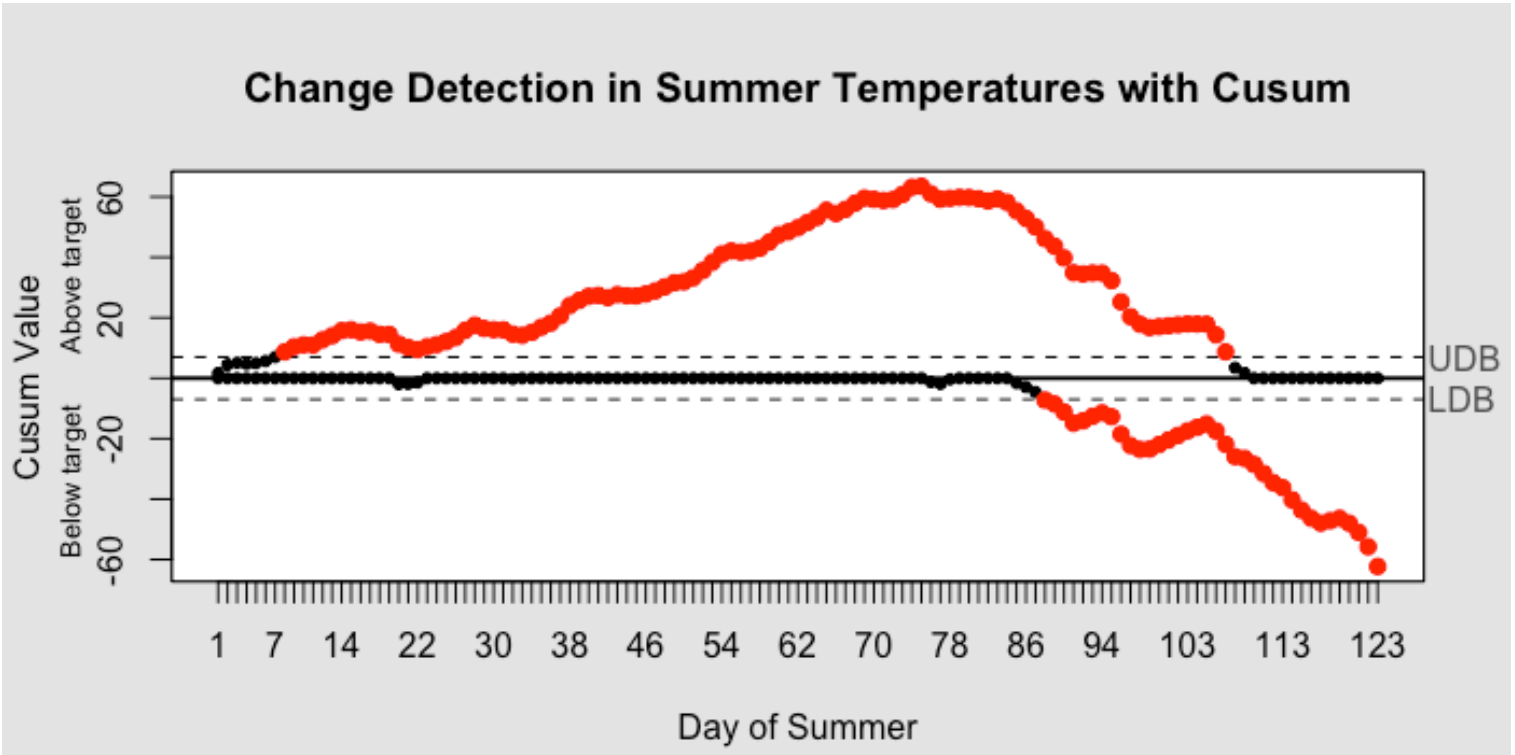
Change Detection in Summer Temperatures with Cusum





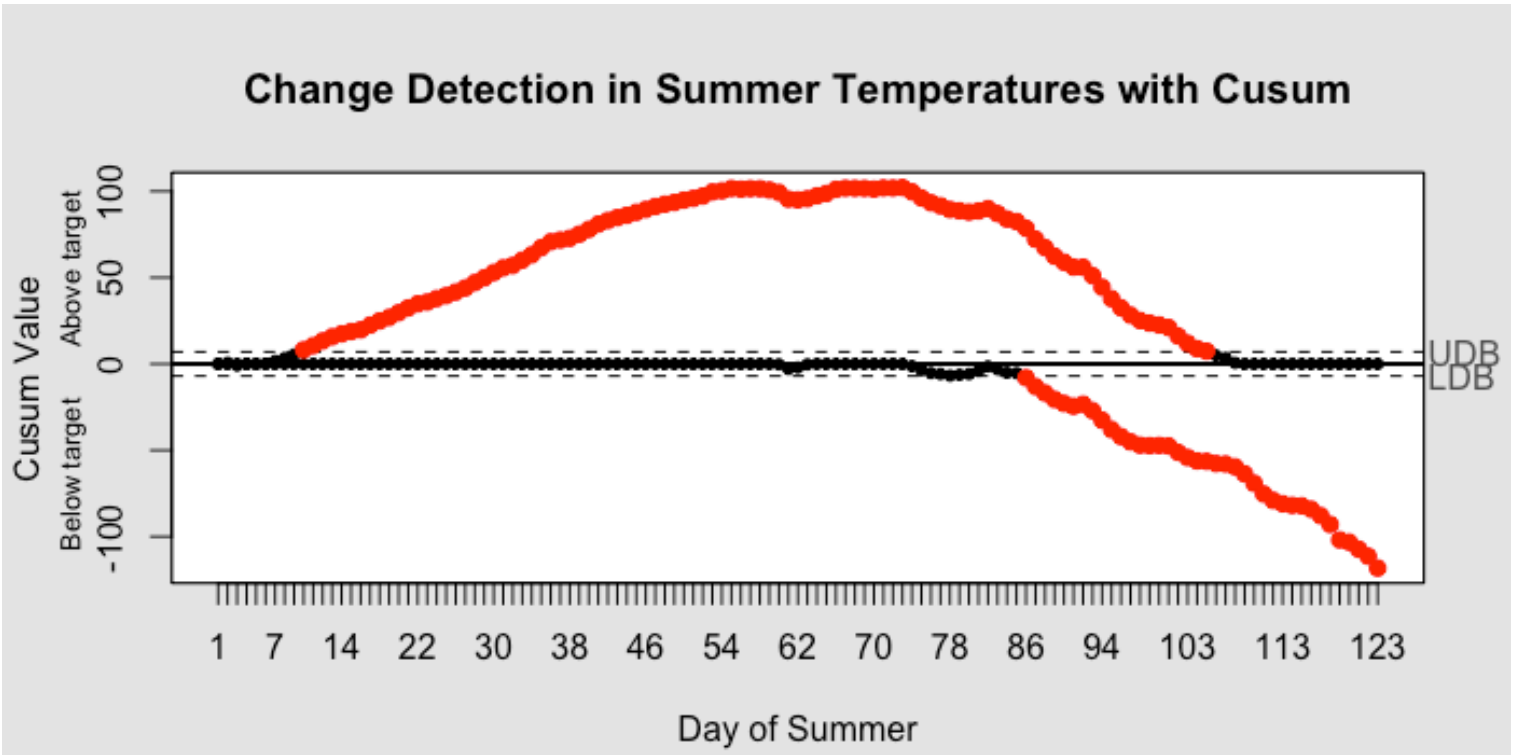
Number of groups = 123
Center = 83.40813
StdDev = 3.299035

Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 77



Number of groups = 123
Center = 83.40813
StdDev = 2.870306

Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 136



Number of groups = 123
Center = 83.40813
StdDev = 2.870306

Decision interval (std. err.) = 7
Shift detection (std. err.) = 1.2
No. of points beyond boundaries = 134

From these charts it is tough to say definitively that the climate is getting warmer. I would lean towards saying yes, particularly because about 7 of the last 10 summers appeared to

be very warm. However, 10 is a rather small sample size and a few recent summers like 2013 seemed very mild. A subject like this warrants further investigation. For the sake of answering the question I would say yes the climate is getting warmer.