WEEK 2 HOMEWORK – SAMPLE SOLUTIONS

IMPORTANT NOTE

These homework solutions show multiple approaches and some optional extensions for most of the questions in the assignment. You don't need to submit all this in your assignments; they're included here just to help you learn more – because remember, the main goal of the homework assignments, and of the entire course, is to help you learn as much as you can, and develop your analytics skills as much as possible!

Question 1

Describe a situation or problem from your job, everyday life, current events, etc., for which a clustering model would be appropriate. List some (up to 5) predictors that you might use.

Here's one answer.

An investor who wants to diversify a portfolio might want to cluster stocks, and then make sure the portfolio does not have too much money invested in any particular cluster.

A common way of clustering is to just classify each company by economic sector or size, but there might be deeper similarities that aren't captured by those factors. So, the investor might create factors related to each stock's performance (such as percent increase/decrease in price) in each quarter over the past 5 years, or each stock's performance in certain key days or intervals, etc. Stocks that behaved similarly would be clustered together.

Question 2

The iris data set contains 150 data points, each with four predictor variables and one categorical response. The predictors are the width and length of the sepal and petal of flowers and the response is the type of flower. The data is available from the R library datasets and can be accessed with iris once the library is loaded. It is also available at the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Iris). The response values are only given to see how well a specific method performed and should not be used to build the model.

Use the R function kmeans to cluster the points as well as possible. Report the best combination of predictors, your suggested value of k, and how well your best clustering predicts flower type.

Here's one possible solution. Please note that a good solution doesn't have to try all of the possibilities in

the code; they're shown to help you learn, but they're not necessary.

The R code in file HW2-Q2.R shows clustering solutions for k=2,3,4,5 using all factors, for both unscaled and scaled data.

	Unscaled data				Scaled data			
	Cluster	Setosa	Versi-color	Virgin-ica	Cluster	Setosa	Versi-color	Virgin-ica
k=2	1	50	3	0	1	50	0	0
	2	0	47	50	2	0	50	50
k=3	1	50	0	0	1	50	0	0
	2	0	48	14	2	0	47	14
	3	0	2	36	3	0	3	36
k=4	1	50	0	0	1	50	0	0
	2	0	27	1	2	0	27	2
	3	0	0	32	3	0	0	29
	4	0	23	17	4	0	23	19
k=5	1	50	0	0	1	28	0	0
	2	0	24	1	2	22	0	0
	3	0	0	24	3	0	27	2
	4	0	0	12	4	0	0	29
	5	0	26	13	5	0	23	19

Table 1. Results using all factors

For k=2, the setosa species is almost perfectly in one cluster, and the other two species (versicolor and virginica) are in the other cluster. For k=3,4,5, setosa is a perfect cluster. When k=4,5 there's a nice cluster of versicolor, a nice cluster or two of virginica, and a cluster of about 40 points that is mixed between the two. k=3 is a little more ambiguous – so even though there are 3 species, it turns out that k=4,5 work better.

The R code also shows clustering solutions for k=2,3,4,5 using only the Petal Length and Petal Width factors, for both unscaled and scaled data.

	Unscaled data				Scaled data			
	Cluster	Setosa	Versi-color	Virgin-ica	Cluster	Setosa	Versi-color	Virgin-ica
k=2	1	50	1	0	1	50	0	0
	2	0	49	50	2	0	50	50
k=3	1	50	0	0	1	50	0	0
	2	0	48	4	2	0	48	4
	3	0	2	46	3	0	2	46
k=4	1	50	0	0	1	50	0	0
	2	0	26	0	2	0	42	0
	3	0	0	35	3	0	0	27
	4	0	24	15	4	0	8	23
k=5	1	50	0	0	1	50	0	0

	2	0	22	0	2	0	23	0
	3	0	0	30	3	0	25	4
	4	0	0	13	4	0	0	25
	5	0	28	7	5	0	2	21

Table 2. Results using only Petal Length and Petal Width factors

Using only the Petal Length and Petal Width factors significantly improves the k=3 solution, and the k=5 solution. Notice that for k=4 especially, using scaled data is a big improvement over using unscaled data.

The R code also introduces the ggplot2 library for plotting, just for your learning pleasure – it's not required for the assignment.

Of course, we can only create the tables above because we happen to know the correct species for each data point. Normally when we're doing clustering, we don't have that information. Instead, we can look at a measure like the total distance between points and their cluster centers in each clustering solution, as shown in the elbow diagram below for scaled data using only the petal factors.

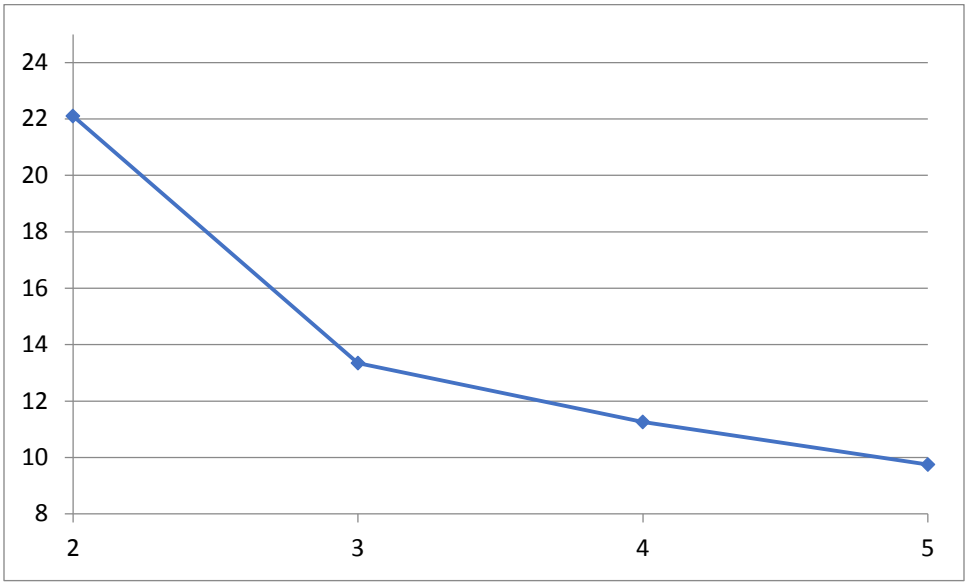

Figure 1. Elbow diagram for scaled data using only petal factors.

Based on this figure, the 3-cluster solution might be the one we would recommend, since k=3 is where the improvements level out.

Question 3

Using crime data from http://www.statsci.org/data/general/uscrime.txt (description at http://www.statsci.org/data/general/uscrime.html), test to see whether there is an outlier in the last column (number of crimes per 100,000 people). Is the lowest-crime city an outlier? Is the highest-crime city an outlier? Use the grubbs.test function in the outliers package in R.

Here's one possible solution. Please note that a good solution doesn't have to try all of the possibilities in the code; they're shown to help you learn, but they're not necessary.

The file HW2-Q3.R contains R code and some explanation for the following approach.

First, because the Grubbs test assumes normality, we start by running a normality test that you'll probably remember from basic statistics: the Shapiro-Wilk test. The test actually suggests that the data is not normally distributed (p=0.001882) – but looking at the Q-Q plot below, it seems that the reason for the non-normality is the tails, which might imply that the test is affected by potential outliers. The middle of the distribution looks normal, so we'll go ahead with the Grubbs test.

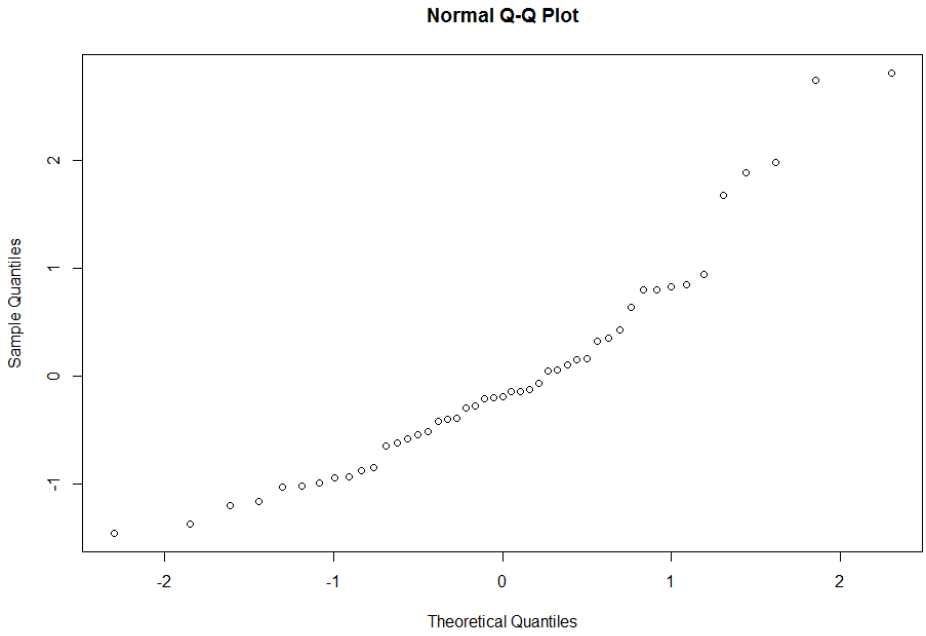

Figure 2. Q-Q plot of the Crime column.

Note here that this is really a judgment call. On the one hand, it could be that the Shapiro-Wilk test is identifying that the tails, especially on the upper end, are really not normally-distributed, enough so that the extreme values aren't really outliers, they're just part of the distribution. On the other hand, it could be that the distribution really is close enough to normal, and the reason it fails the Shapiro-Wilk test is that there's outlying data. The Grubbs test's validity depends on which of these is closer to true.

In this case, let's go on with the Grubbs test. At worst, it'll either show that there aren't outliers, or it'll identify potential outliers – then we would (if this was more than a homework assignment) investigate those data points more carefully to see what's going on, to determine whether they seem like a real part of the distribution or whether they're real outliers.

It turns out that the lowest-crime city is unlikely to be an outlier (p-value so close to 1 that it just comes up as 1).

On the other hand, the highest-crime city might be an outlier (p=0.079), and if we remove it, the second-highest-crime city also appears to be an outlier (p=0.028). The box-and-whisker plot below shows the outliers more clearly.

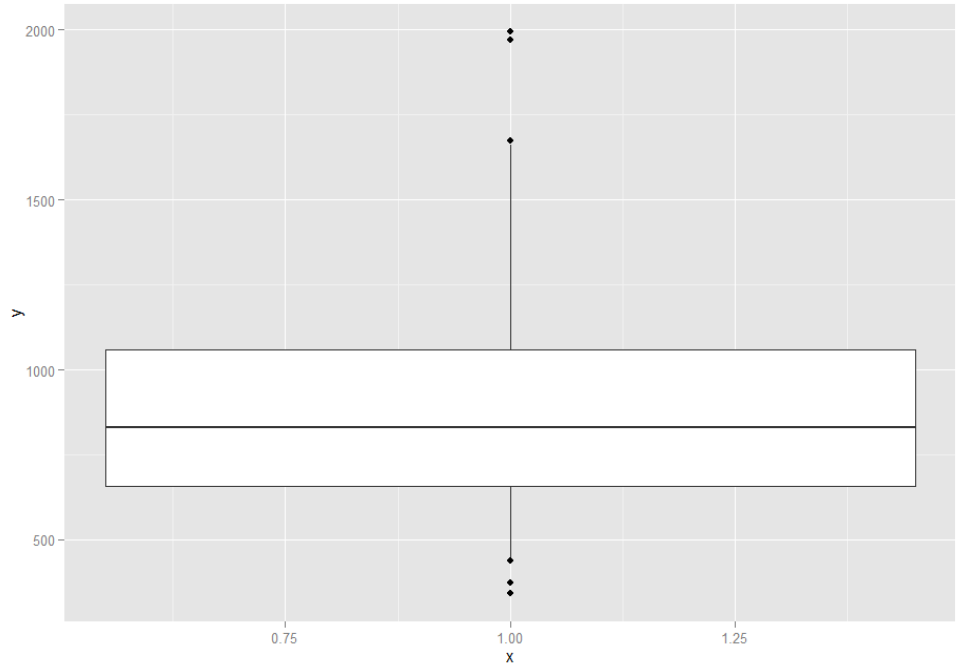

Figure 3. Box-and-whisker plot of the Crime column.

Note that some people tried to determine whether there was an outlier on the high end and on the low end simultaneously, using the "type=11" parameter in the grubbs.test() R function. The problem with this approach is that the answer it returns is "no" – because they're not *both* outliers (as we saw, the lowest-crime city isn't an outlier). That result hides the fact that the highest-crime city probably *is* an outlier (and in fact, so is the second-highest). So using the "type=10" parameter is generally a better approach; it tests one side, and adding the "opposite=TRUE" parameter tests the other side. See HW2-Q3.R for details.

Question 4

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?

Here's one answer.

Bird flu is a common disease in certain parts of East Asia, Middle East and West Africa. Public health organizations (such as CDC) want to identify a potential outbreak as soon as possible so that action can be taken to stop its spread. CUSUM can be used to monitor the number of cases of this disease, and detect a change when the number of cases rises above a threshold, indicating a possible outbreak.

Ideally, the CUSUM statistic S_t should remain smaller than the threshold if an epidemic is not going to occur, and quickly cross the threshold if an epidemic is going to occur. In some uses of CUSUM, a typical choice for threshold T is 5 standard deviations, while C will be half of a standard deviation. But in the

case of a potentially-deadly epidemic, those "standard" values are probably too conservative. The cost of a false alarm is much lower than the cost of waiting too long to detect a change. Therefore, to save lives, the values of both T and C should be lower (and can be calibrated based on previous data).

Question 5

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. That involves finding a good critical value and threshold to use across all years. You can get the data that you need 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 straightforward enough that an Excel spreadsheet can easily do the job too.*

Here are two approaches. Please note that there are other good approaches too, and a good solution doesn't have to try all of the possibilities in here; they're shown to help you learn, but they're not necessary. It also might be that some people interpreted the question differently, and have a good solution to the question as they understood it – that's perfectly fine too. All I really care about is that you're learning how to use the CUSUM model.

The file HW2-Q5.xlsx (sheet Part 1 Approach 1) shows the first approach. The "temps" sheet shows the raw data, and the "Part 1 Approach 1" sheet shows a CUSUM calculation. (Note: Column C is set to turn green starting at the first detected change, using conditional formatting.) For each day (July 1 through October 31), we take the average temperature over the 20-year period 1996-2015. To define μ, the average temperature before any change, we use the average of the average July temperatures (i.e., the average of: July 1's average, July 2's average, …, July 31's average)[1]. That turns out to be 88.75, which is pretty reasonable compared to the July average temperature range from 87 to 90.4.

By changing the values in cells D1 and F1, you can test different values of C and T, some of which are shown in this chart:

C	T	Change detected	Avg temp on change date
5	25	Sep 26	78.55
5	15	Sep 24	80.4
5	5	Sep 18	81.7
1	25	Sep 9	85.05
1	15	Sep 6	84.6
1	5	Sep 1	85.25
0.44262	4.42625	Aug 30	85.8

Table 3. Change detection for different values of C and T. The last row shows C as ½ standard deviation of the July data, and T as 5 standard deviations of the July data, which is sometimes used as a default.

[1] Of course, this is the same as the average over all 620 July observations (31 days each year for 20 years).

My personal judgment (which you don't have to agree with – there's not a single "right" answer) is that the default values are too sensitive. There is a small change (from the 87-88 range to the 85 range) that starts on August 30, but that change seems too small to call the end of summer. Something closer to Sep 24 or 25 seems like a more-significant change. On the other hand, the graph below does seem to show a change around the end of August (close to Labor Day, the traditional end of summer).

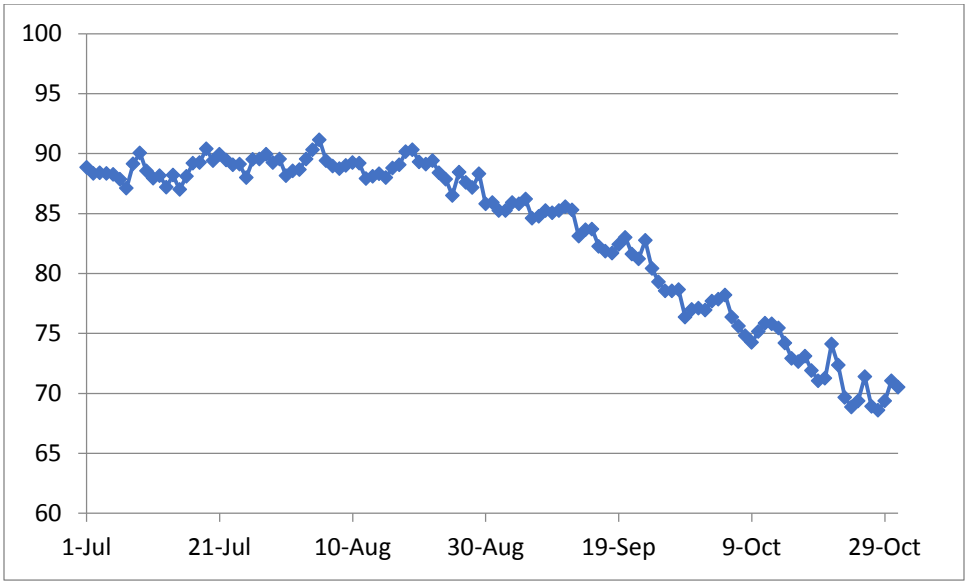

Figure 4. Average (1996-2015) temperature day by day.

A second approach is shown in the file HW2-Q5.xlsx (sheet "Part 1 Approach 2"). In this approach, we find the change point for each year individually, and then average them.

In the "CUSUM" sheet, each column B through U corresponds to a year. B12:B134, C12:C134, etc. are the values of S_t for 1996, 1997, etc. Rows 1-4 show the July average, max, min, and standard deviation each year. Rows 5 through 7 show the values of C, T, and mu each year. Rows 8 and 9 show the date a change is detected each year. Finally, cells V8 and V9 show the average day that change is detected.

As in the first approach, the key is finding the right values of C and T. Looking at the year-to-year data, there are certainly some days where the temperature is temporarily 10-15 degrees colder than average, but then goes back up – we don't want to mistakenly identify that randomness as a change.

C	**T**	**Avg date change detected**	**Avg temp on change date**
0	35	Aug 18	90.15
5	35	Sep 19	82.4
10	35	Oct 7	75.6
0	25	Jul 22	89.45
5	25	Sep 17	79.75
10	25	Oct 4	77.85
0	15	Jul 17	89.2

5	15	Sep 2	85.25
10	15	Sep 26	78.55

Table 4. End-of-summer estimates using approach 2.

As above, this is a judgment call, but my preference would be for C=5 and T=25 or 35, giving a date around Sep 17 or Sep 19. Figure 5 below shows the year-by-year data.

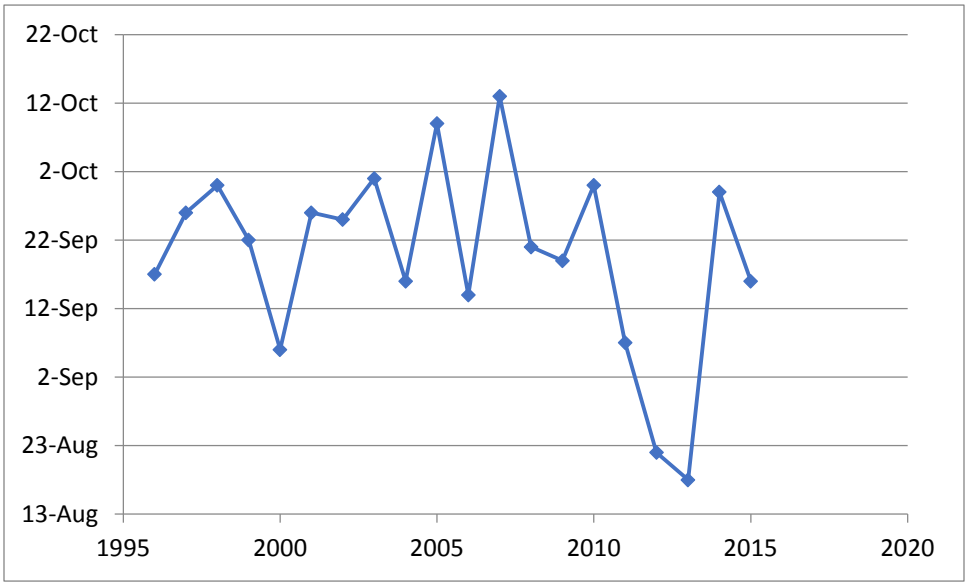

Figure 5. Year-by-year end-of-summer estimates.

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).*

Again, here are two approaches. Please note that there are other good approaches too, and a good solution doesn't have to try all of the possibilities in here; they're shown to help you learn, but they're not necessary. It also might be that some people interpreted the question differently, and have a good solution to the question as they understood it – that's perfectly fine too. All I really care about is that you're learning how to use the CUSUM model.

One approach is to take the yearly end-of-summer date estimate from Part 1 Approach 2, and see if there's a change over time. In other words, "is summer lasting longer than it did before?" This is shown in HW2-Q5.xlsx (sheet "Part 2 Approach 1"). The values of C and T we choose make the difference between a "yes" or "no" answer.

C	**T**	**Year change detected**	**Change detected in year**
3	60	Never	n/a
3	45	2007	Oct 13
3	30	2005	Oct 9
3	15	1999	Sep 22

7	60	Never	n/a
7	45	Never	n/a
7	30	Never	n/a
7	15	2007	Oct 13
14	60	Never	n/a
14	45	Never	n/a
14	30	Never	n/a
14	15	Never	n/a

Table 5. End-of-summer estimates.

As you can see, this is ambiguous. It's not clear if summer is ending later. In fact, as Figure 6 shows below, if anything there was a blip in 2012-2013 where our Part 1 Approach 2 model shows summer ending significantly *earlier*.

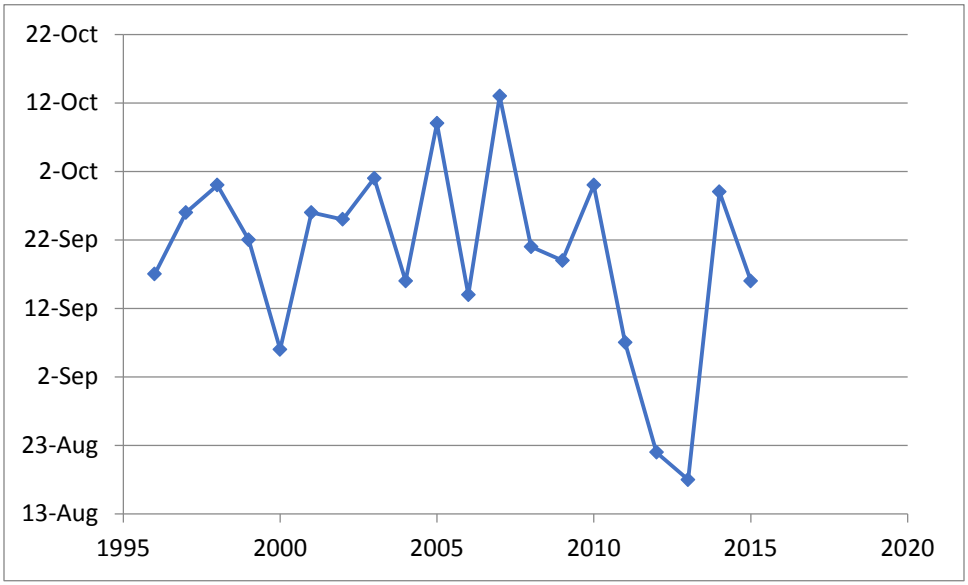

Figure 6. Year-by-year end-of-summer estimates (same as Figure 5).

A second approach is to answer "however long summer is, are those days getting hotter?" As shown in HW2-Q5.xlsx (sheet "Part 2 Approach 2", for each year we could take the end-of-summer estimate from Part 1 Approach 2, and calculate the average daily temperature through that date.

Figure 7 shows the year-by-year average summer temperature estimates. Visually, it seems difficult to identify a change point – 2010-2012 appears higher, but then the 2013-2015 drop back to the standard range. CUSUM similarly is ambiguous about whether/when there was a change; see Table 6. For more-sensitive values of C and T, change is detected in 2011 or 2012, but other values of C and T don't detect a change, because the temperatures drop back starting in 2013.

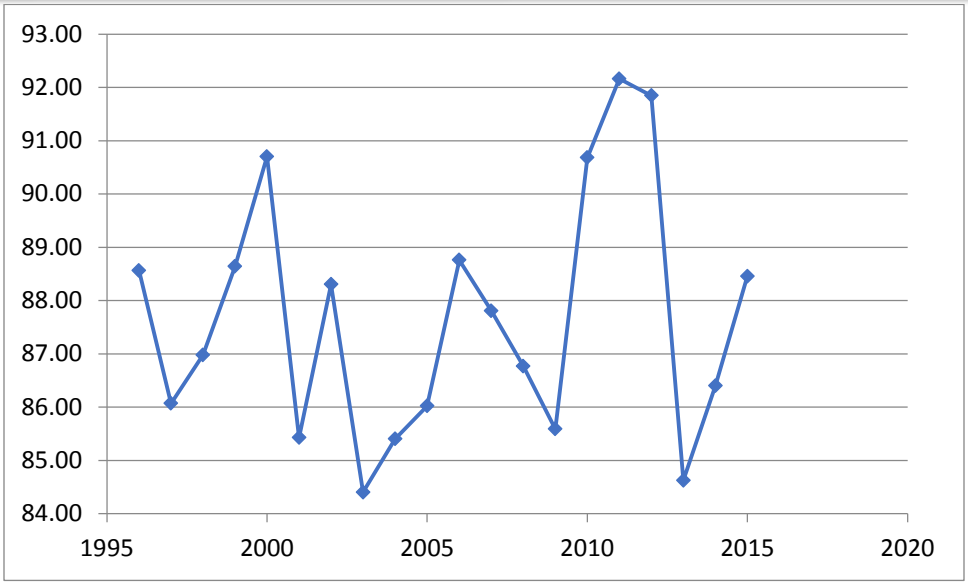

Figure 7. Year-by-year average summer temperature estimates.

C	T	Year change detected	Avg temp through day
0	5	2011	92.16
0	10	Never	n/a
0	15	Never	n/a
1	5	2012	91.85
1	10	Never	n/a
1	15	Never	n/a
2	5	Never	n/a
2	10	Never	n/a
2	15	Never	n/a

Table 6. End-of-summer estimates using approach 2.