hw_4

Jeff Tilton

9/15/2018

Question 7.1

Describe a situation or problem from your job, everyday life, current events, etc., for which exponential smoothing would be appropriate. What data would you need? Would you expect the value of α (the first smoothing parameter) to be closer to 0 or 1, and why?

Response

Time series data have inherent randomness that can make it difficult to understand the underlying signal. An observer has to consider if a new value that is different than the baseline data is a true change or just a random event. Exponential smoothing is a method that balances these considerations in 2 ways.

1. We may consider the observed value as a true change in the baseline. Therefore:

$$S_t = x_t$$

2. Or we might think that there is no change to the baseline and that the observed change is just due to randomness. Therefore:

$$S_t = S_{t-1}$$

Exponential smoothing combines the two with the α coefficient.

$$S_t = \alpha X_t + (1 - \alpha)S_{t-1}$$

The coefficient is used to determine which is more likely, the change is a true change ($S_t = x_t$) then α will be closer to 1. Or the change is due to randomness ($S_t = S_{t-1}$) α is closer to 0.

This modeling technique can be used in any time series problem. I work as an hydraulic engineer in reservoir control office. Dams often need to spill to decrease fish mortality. Spill, however, increases the total dissolved gas (TDG) in the river, which also increases fish mortality. Gages that read TDG values are noisy due to the turbulence spill creates. Exponential smoothing can be used to create a better understanding of the baseline.

Question 7.2

Using the 20 years of daily high temperature data for Atlanta (July through October) from Question 6.2 (file temps.txt), build and use an exponential smoothing model to help make a judgment of whether the unofficial end of summer has gotten later over the 20 years. (Part of the point of this assignment is for you to think about how you might use exponential smoothing to answer this question. Feel free to combine it with other models if you'd like to. There's certainly more than one reasonable approach.)

Note: in R, you can use either HoltWinters (simpler to use) or the smooth package's es function (harder to use, but more general). If you use es, the Holt-Winters model uses model="AAM" in the function call (the first and second constants are used "A"dditively, and the third (seasonality) is used "M"ultiplicatively; the documentation doesn't make that clear).

Response

Data Visualization

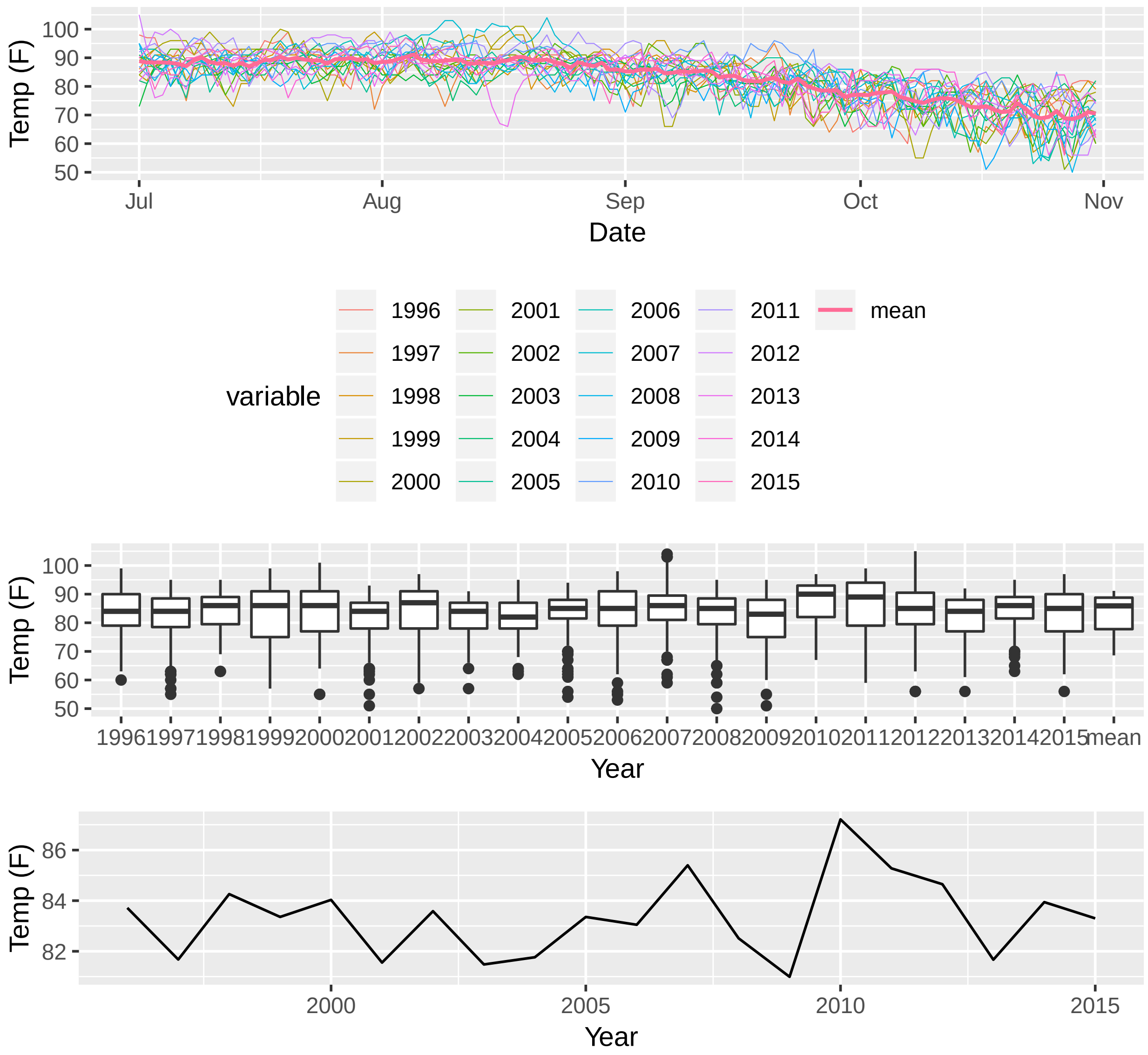

Figure 1: Atlanta (July through October) daily high temperature data

Figure 2 From top to bottom: (1) Daily temperature Atlanta highs from 1996 to 2015 as well as mean daily highs. (2) Atalanta yearly spread of daily high temperatures from July - October. (3) Atalanta yearly mean temperature.

The mean daily temperature highs, seen above, is relatively stable. Any given year has a considerable amount of randomness, which can be seen in the daily temperature highs and the yearly spread. There are many values below the 75th qualntile in the majority of years seen in both the boxplots and the downward spikes in plot 1.

The unofficial end of summer, defined as when the weather starts to cool off each summer in Atlanta occurs sometime in mid August. The mean daily temperature highs dip within that timeframe.

Goals

This exercise has 1 goal:

1. Make a judgment of whether the unofficial end of summer has gotten later over the 20 years

Method

1. Create smoothed series for each year using Holt-Winters
2. Use Cusum on smoothed series to determine when official summer ends for each year
3. Collect th first date where $S_t >= T$
4. See if official end of summer has gotten later

Assumptions

I assume daily temperature values to be random throughout the year. Therefore I will use higher c and t values for the CUSUM algorithm.

CUSUM

$$S_t = max\{0, S_{t-1} + (\mu - X_t - C)\}$$

I will set $c = 4$ and $t = 7$. I will take the mean of all years through August 10th for μ.

Results

The below table is the first day in each year when $S_t >= T$.

year	days
1996	2-Sep
1997	7-Jul
1998	11-Sep
1999	13-Jul
2000	25-Jul
2001	31-Aug
2002	12-Jul
2003	6-Jul
2004	10-Aug
2005	7-Jul
2006	8-Jul
2007	16-Sep
2008	25-Aug
2009	20-Jul
2010	27-Sep
2011	5-Sep
2012	30-Aug
2013	3-Jul
2014	20-Jul
2015	30-Aug
mean	17-Sep

Discussion

I am not sure if the method I have chosen is extremely robust. I feel like I have been cherry picking the c and t values to get something that I want. The above table does not seem to indicate that the unofficial end of summer is getting later in the year, but tends to jump around. I look forward to seeing other solutions.

Just for Fun

Goals

This exercise has 1 goal:

1. Determine if summers are getting hotter

Method

1. Run the CUSUM method on the yearly mean and median time series
2. Use the Holt-Winters function on each year to obtain individual α values
3. Use the Holt-Winters function on the yearly mean and median time series to obtain individual α values

Assumptions

I assume yearly mean and median values to be stable and even small increases in temperature are significant. Therefore I will use low c and t values for the CUSUM algorithm. If the CUSUM algorithm detects a change at a given year, and that year has a low α value then the change is not random and I can say that temperatures are increasing. The α value is being used as a proxy for randomness.

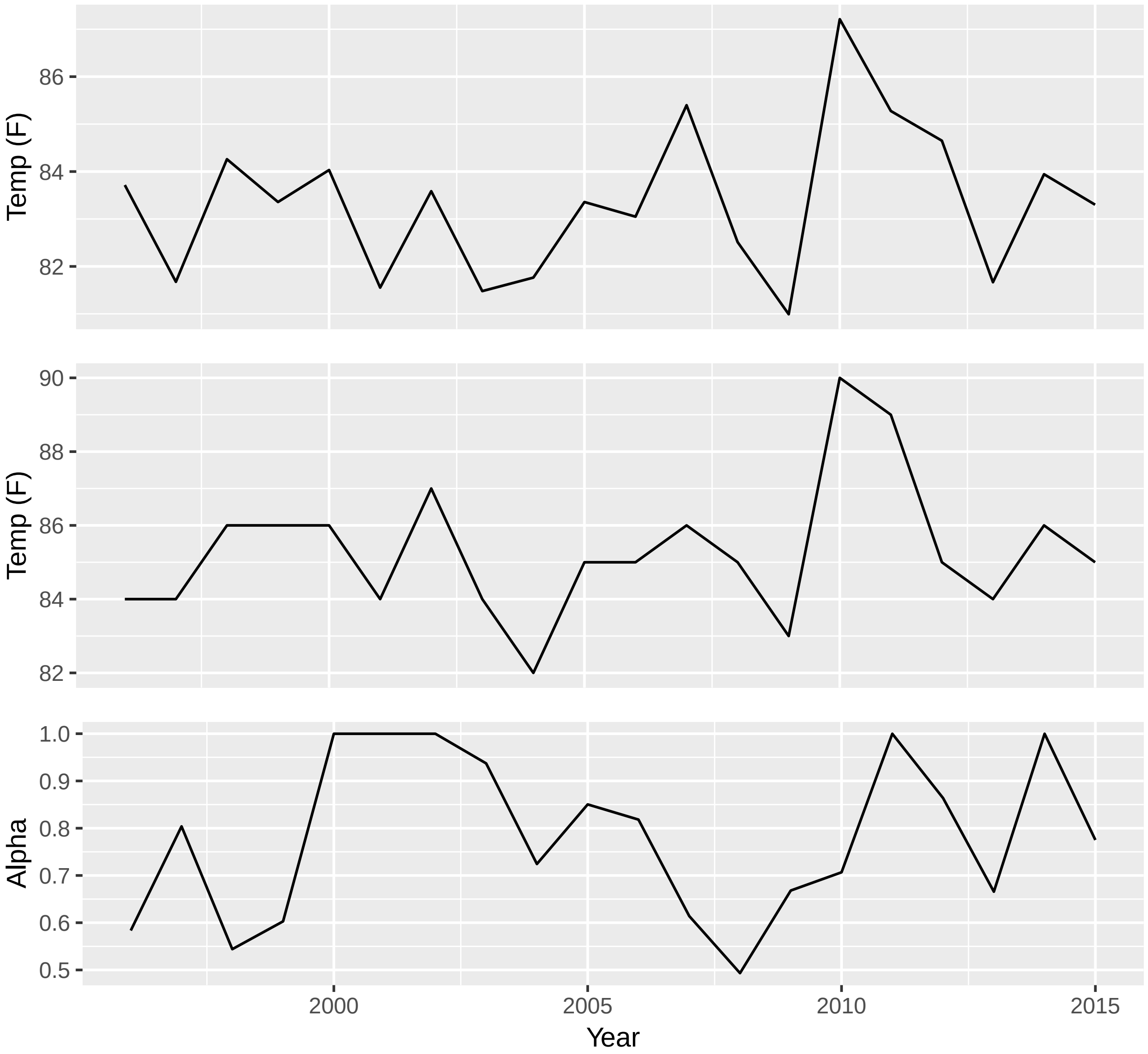

Figure 2: Atlanta (July through October) Mean, median daily high temperatures and yearly alpha values for Holt-Winters

Figure 2 From top to bottom: (1) Yearly mean Atlanta daily high temperatures from July - October. (2) Yearly mean Atlanta daily high temperatures from July - October. (3) Yearly α value results from Holt-Winters.

The above plots suggest that on the coldest years, example 2008, are due to randomness because the corresponding α value is low. Where the hottest year, 2011, has an α near 1, meaning the values are not random.

CUSUM

$$S_t = max\{0, S_{t-1} + (\mu - X_t - C)\}$$

I will set $c = 2$ and $t = 1$. I will use the series' mean value for μ.

Table 1: Results of the CUSUM algorithm on Atlanta yearly mean and median daily high temperature with $c = 2$ and $t = 1$

year	mean	median
1996	FALSE	FALSE
1997	FALSE	FALSE
1998	FALSE	FALSE
1999	FALSE	FALSE
2000	FALSE	FALSE
2001	FALSE	FALSE
2002	FALSE	FALSE
2003	FALSE	FALSE
2004	FALSE	FALSE
2005	FALSE	FALSE
2006	FALSE	FALSE
2007	FALSE	FALSE
2008	FALSE	FALSE
2009	FALSE	FALSE
2010	TRUE	TRUE
2011	TRUE	TRUE
2012	TRUE	TRUE
2013	FALSE	FALSE
2014	FALSE	FALSE
2015	FALSE	FALSE

The results from the CUSUM algorithm suggest that there is a temperature change starting in year 2010 and continuing through 2012 for both the mean and median series. The corresponding α values are relatively high, between .7 and 1. The remaining 3 years, 2013-2015 did not show a change.

The corresponding $\alpha's$ for the mean and median series are 0.0041 and 0.14. Values this low indicate there is a significant amount of randomness in the data.

Discussion

I cannot say with any amount of certainty that daily temperatures are increasing based on the above results. Although the CUSUM algorithm detected a change for a few years, the change did not persist. Furthermore the yearly mean and median series shows a significant amount of randomness that is not accounted for in this simplistic model.