Week 3 Assignment - Time Series Analysis with Exponential Smoothing

Omer Farooq (EDx ID: mfarooq4)

02/04/2020

Table of Contents

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?

At my job at the T-Mobile HQ in the Seattle area, my team helps get analytics products built for our network supply chain team. This team manages the planning, procurement and logistics of getting the right equipment to the right locations so that T-Mobile's network could get built or improved. One of the things that would be interesting to capture and apply exponential smoothing model on would be the cycle time of picking a kit. This is the time it takes the distribution center to completely pick, pack and stage a kit of materials after the order has been placed in the Warehouse Management System. Our supply chain team uses an average of how long this cycle time typically is and a thorough time series analysis would help improve the benchmark for this cycle time by incorporating trend and seasonability. For example, it would be appropriate to hold the distribution center team accountable to a different benchmark during the busy season.

Collecting a daily average cycle time for a period of 1 year would give enough time series data to do the analysis on. Given we do have a busy season during summer when more network upgrades happen, I would expect the exponential smoothing model to show seasonality in the data. It would be interesting to see if any trends come up in the model as well.

As for the value of α, as it is a measure of randomness in the data, I don't expect the cycle times to vary a lot day over day other than may be seasonability or any trend affects. Thus, a higher value of α, closer to 1 but possibly between 0.5 and 1 would be a expected.

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).

First, I loaded the needed libraries.

```
library(ggplot2)
library(dplyr)
library(timetk)
library(forecast)
library('tseries')
```

Next, I set the seed for consistent results and loaded the temps data. Also, ran a quick summary of the data to check basic statistics of each column.

```
#setting the seed so that results are the same at every run
set.seed(101)

#loading data
temps <- read.delim("data_7.2/temps.txt")

#quick glance at the data
head(temps)

##     DAY X1996 X1997 X1998 X1999 X2000 X2001 X2002 X2003 X2004 X2005 X2006
X2007
## 1 1-Jul    98    86    91    84    89    84    90    73    82    91    93
95
## 2 2-Jul    97    90    88    82    91    87    90    81    81    89    93
85
## 3 3-Jul    97    93    91    87    93    87    87    87    86    86    93
82
## 4 4-Jul    90    91    91    88    95    84    89    86    88    86    91
86
## 5 5-Jul    89    84    91    90    96    86    93    80    90    89    90
88
## 6 6-Jul    93    84    89    91    96    87    93    84    90    82    81
87
```

```
##   X2008 X2009 X2010 X2011 X2012 X2013 X2014 X2015
## 1    85    95    87    92   105    82    90    85
## 2    87    90    84    94    93    85    93    87
## 3    91    89    83    95    99    76    87    79
## 4    90    91    85    92    98    77    84    85
## 5    88    80    88    90   100    83    86    84
## 6    82    87    89    90    98    83    87    84
```

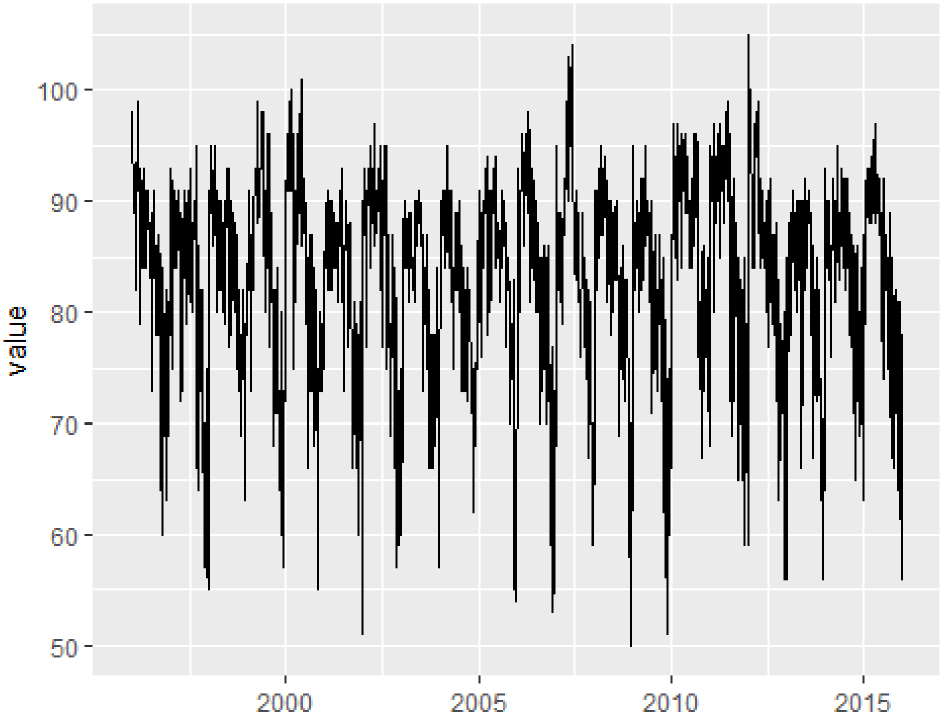
```
#basic stats of the temps data
summary(temps)
```

```
##       DAY           X1996           X1997           X1998           X1999      
##  1-Aug  :  1   Min.   :60.00   Min.   :55.00   Min.   :63.00   Min.   :57.00  
##  1-Jul  :  1   1st Qu.:79.00   1st Qu.:78.50   1st Qu.:79.50   1st Qu.:75.00  
##  1-Oct  :  1   Median :84.00   Median :84.00   Median :86.00   Median :86.00  
##  1-Sep  :  1   Mean   :83.72   Mean   :81.67   Mean   :84.26   Mean   :83.36  
##  10-Aug :  1   3rd Qu.:90.00   3rd Qu.:88.50   3rd Qu.:89.00   3rd Qu.:91.00  
##  10-Jul :  1   Max.   :99.00   Max.   :95.00   Max.   :95.00   Max.   :99.00  
##  (Other):117                                                                  
##      X2000            X2001           X2002           X2003      
##  Min.   : 55.00   Min.   :51.00   Min.   :57.00   Min.   :57.00  
##  1st Qu.: 77.00   1st Qu.:78.00   1st Qu.:78.00   1st Qu.:78.00  
##  Median : 86.00   Median :84.00   Median :87.00   Median :84.00  
##  Mean   : 84.03   Mean   :81.55   Mean   :83.59   Mean   :81.48  
##  3rd Qu.: 91.00   3rd Qu.:87.00   3rd Qu.:91.00   3rd Qu.:87.00  
##  Max.   :101.00   Max.   :93.00   Max.   :97.00   Max.   :91.00  
##                                                                  
##      X2004           X2005           X2006           X2007      
##  Min.   :62.00   Min.   :54.00   Min.   :53.00   Min.   : 59.0  
##  1st Qu.:78.00   1st Qu.:81.50   1st Qu.:79.00   1st Qu.: 81.0  
##  Median :82.00   Median :85.00   Median :85.00   Median : 86.0  
##  Mean   :81.76   Mean   :83.36   Mean   :83.05   Mean   : 85.4  
##  3rd Qu.:87.00   3rd Qu.:88.00   3rd Qu.:91.00   3rd Qu.: 89.5  
##  Max.   :95.00   Max.   :94.00   Max.   :98.00   Max.   :104.0  
##                                                                 
##      X2008           X2009           X2010           X2011      
##  Min.   :50.00   Min.   :51.00   Min.   :67.00   Min.   :59.00  
##  1st Qu.:79.50   1st Qu.:75.00   1st Qu.:82.00   1st Qu.:79.00  
##  Median :85.00   Median :83.00   Median :90.00   Median :89.00  
##  Mean   :82.51   Mean   :80.99   Mean   :87.21   Mean   :85.28  
##  3rd Qu.:88.50   3rd Qu.:88.00   3rd Qu.:93.00   3rd Qu.:94.00  
##  Max.   :95.00   Max.   :95.00   Max.   :97.00   Max.   :99.00  
##                                                                 
##      X2012            X2013           X2014           X2015     
##  Min.   : 56.00   Min.   :56.00   Min.   :63.00   Min.   :56.0  
##  1st Qu.: 79.50   1st Qu.:77.00   1st Qu.:81.50   1st Qu.:77.0  
##  Median : 85.00   Median :84.00   Median :86.00   Median :85.0  
##  Mean   : 84.65   Mean   :81.67   Mean   :83.94   Mean   :83.3  
##  3rd Qu.: 90.50   3rd Qu.:88.00   3rd Qu.:89.00   3rd Qu.:90.0  
##  Max.   :105.00   Max.   :92.00   Max.   :95.00   Max.   :97.0  
##
```

Then, I started preparing the data to use in the HoltWinters model. HoltWinters model inputs a time series object, so I converted the data to vector first and then a time series object. A basic plot of the temperatures over 20 years was also created. It showed the cyclic nature of the data but given the density of the graph, did not provide clear evidence to answer the questions whether unofficial end of summer gotten later over 20 years. I utilised CUSUM method to analyse this question in the last assignment. In this one, I used the HoltWinters to smooth the data first and then used it's output in for CUSUM analysis to see if summer ended later over the years.

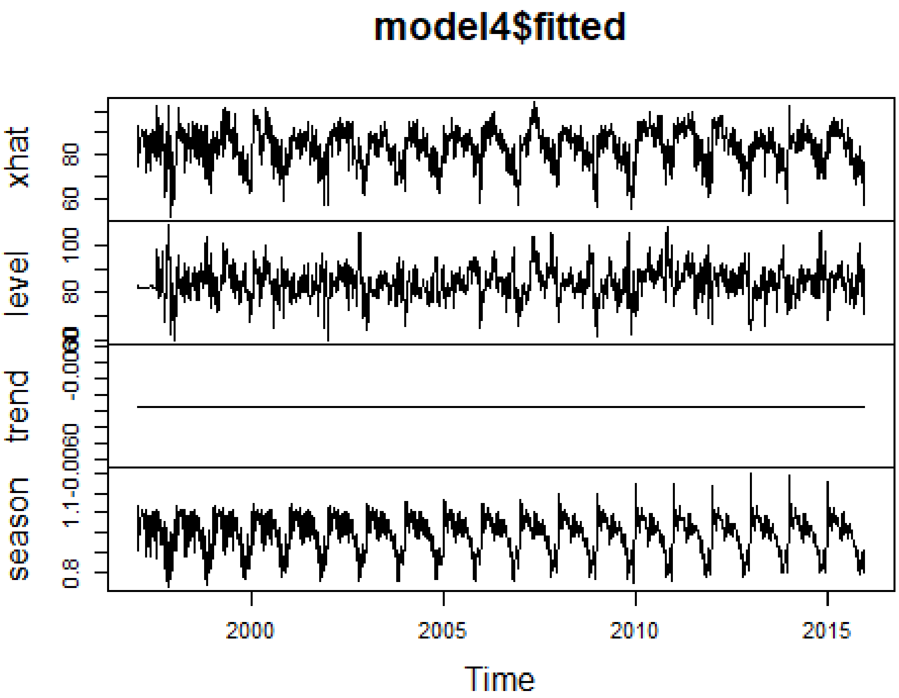
Another difference from the last assignment is that, my interpretation of this questions last time was whether the summer got hotter over the years. For that, I looked at average temperature for July, Aug and Sept over 20 years to see if the average moved up. To check whether summer ended later over the years, I did the CUSUM differently than the last assignment here. I performed CUSUM on each year's smoothed data points and saw if the end of summer moved later over the years.

```r
#creating a vector or yearly temp values
vector_temps <- as.vector(unlist(temps[,2:21]))

#creating a timeseries object
ts_temps <- ts(vector_temps,start = 1996, frequency = nrow(temps))

#plot of timeseries
dataframe <-tk_tbl(ts_temps)
ggplot(dataframe, aes(x = index, y=value)) + geom_line() + xlab("")
```



Next is the key part of the assignment, I applied different Holtwinters models on the time series data and logged the results in a matrix.

```r
#empty matrix to log models resuls
test_results <- matrix(NA, nrow=4, ncol=5)
colnames(test_results) <- c("TEST METHOD","ALPHA (RANDOMNESS)","BETA
(TREND)","GAMMA(SEASONALITY)", "SSE")

#single exponential smoothing
model1 <- HoltWinters(ts_temps,beta=FALSE,gamma=FALSE)
test_results[1,] <- c("Single Exponential Smoothing",
round(model1$alpha,digit=4), model1$beta, model1$gamma, round(model1$SSE))

#Double exponential smoothing
model2 <- HoltWinters(ts_temps,gamma=FALSE)
test_results[2,] <- c("Double Exponential Smoothing",
round(model2$alpha,digit=4), round(model2$beta,digit=4), model2$gamma,
round(model2$SSE))

#Tripple exponential smoothing - additive
model3 <- HoltWinters(ts_temps)
test_results[3,] <- c("Tripple Exponential Smoothing (Additive)",
round(model3$alpha,digit=4), round(model3$beta,digit=4),
round(model3$gamma,digit=4), round(model3$SSE))

#Tripple exponential smoothing - multiplicative
model4 <- HoltWinters(ts_temps,seasonal = "multiplicative")
test_results[4,] <- c("Tripple Exponential Smoothing (Multiplicative)",
round(model4$alpha,digit=4), round(model4$beta,digit=4),
round(model4$gamma,digit=4), round(model4$SSE))

#printing the resuls matrix
test_results

##      TEST METHOD                                      ALPHA (RANDOMNESS)
## [1,] "Single Exponential Smoothing"                   "0.8388"
## [2,] "Double Exponential Smoothing"                   "0.8446"
## [3,] "Tripple Exponential Smoothing (Additive)"       "0.6611"
## [4,] "Tripple Exponential Smoothing (Multiplicative)" "0.615"
##      BETA (TREND) GAMMA(SEASONALITY) SSE
## [1,] "FALSE"      "FALSE"            "56198"
## [2,] "0.0037"     "FALSE"            "56573"
## [3,] "0"          "0.6248"           "66244"
## [4,] "0"          "0.5495"           "68905"

#plotting model4 output
plot(model4$fitted)
```



From the results above, the **Single exponential smoothing** with trend and seasonality parts turned off showed a higher (0.84) α indicating a low randomness in the data. A higher alpha gave more weight to the last reading over the time periods.

The **double exponential smoothing** with randomness and trend parts turned on showed the same high α indicating low randomness and very low (0.004) β indicating lack any trends in the data.

Lastly, **tripple exponential smoothing** with all 3 parts turned on, first for *additive* seasonality and then for *multiplicative* seasonality. It indicated higher α, very low β and a slightly highr γ indicating some seasonality in the data. I also logged the SSE for these models, which moved up slightly more constants were turned on.

Given there was an indication of seasonality, I used the tripple multiplicative smoothing for the rest of the assignment. I also plotted the xhat, level, trend and season outputs from the fitted object of the tripple holtwinters model. I confirmed the interpretations of the fitted object output in a Piazza post as well, it's explained below:

- **xhat** - this is the combined smoothed x value output from the function (using α, β and γ components of the model)
- **level** - this is equivalent to the St baseline value which is smoothed between the current xhat point and the previous xhat point using α only.
- **trend** - this is the trend component estimate built using β component only/.
- **season** - this is the seasonality factor (additive or multiplicative depending on model settings) built based on γ only.

The xhat plot was pretty similar to the actual data, levels was slightly smoothed, trend was flatline because no trend was found and seasonality plot showed consistent pattern over the years.

Next, I exported the trippled multiplicative holtwinters model data for xhat, level and seasonality and performed CUSUM on all 3 types of data points to see what evidence was there to say whether summer ended later over the period of 19 years (Holtwinters output starts from period 2 given there could not be fitted value for the 1st period).

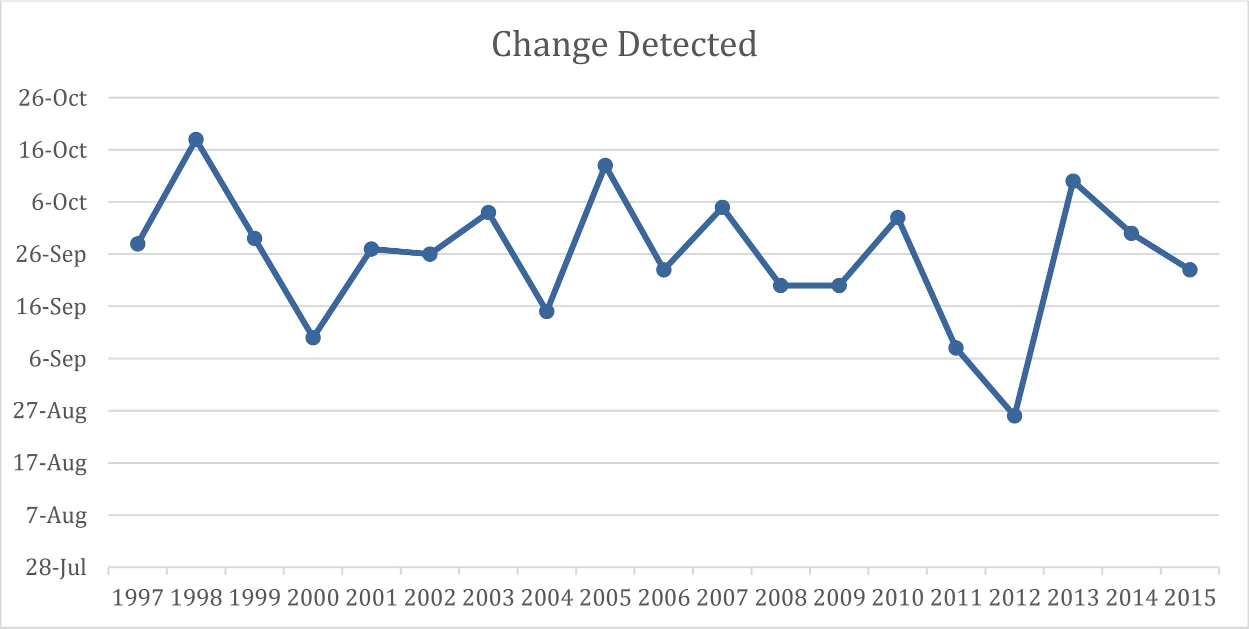
```r
#xhat
x_matrix_model4 <- matrix(model4$fitted[,1], nrow=123)
write.csv(x_matrix_model4, file="7_1_xhat.csv")

#levels
l_matrix_model4 <- matrix(model4$fitted[,2], nrow=123)
write.csv(l_matrix_model4, file="7_1_levels.csv")

#seasonality
s_matrix_model4 <- matrix(model4$fitted[,4], nrow=123)
write.csv(s_matrix_model4, file="7_1_seasonality.csv")
```

Results of CUSUM

The approach I took for this analysis was:

- I applied CUSUM on each year separately on 3 separate datasets i.e. xhat, level, seasonality (outputs of tripple exponential smoothing model).
- I calculated the average (μ) for July averaged temperatures for each year. This makes sense because CUSUM requires a mean for no change data. July is the least change summer month every year.
- I calculated the standard deviation for the same July average temperatures for each year.
- Statistically, C should be 0.5 times standard deviation and T should be 4 times standard deviation. But if the data is too sensitives or less sensitive, higher or lower values might be needed. I checked a range of C and T values to see where change is indicated and selected those for respective datasets. (*source*: https://www.spcforexcel.com/knowledge/variable-control-charts/keeping-process-target-cusum-charts)
- Since we are detecting a decrease, the St formula for decrease is used.

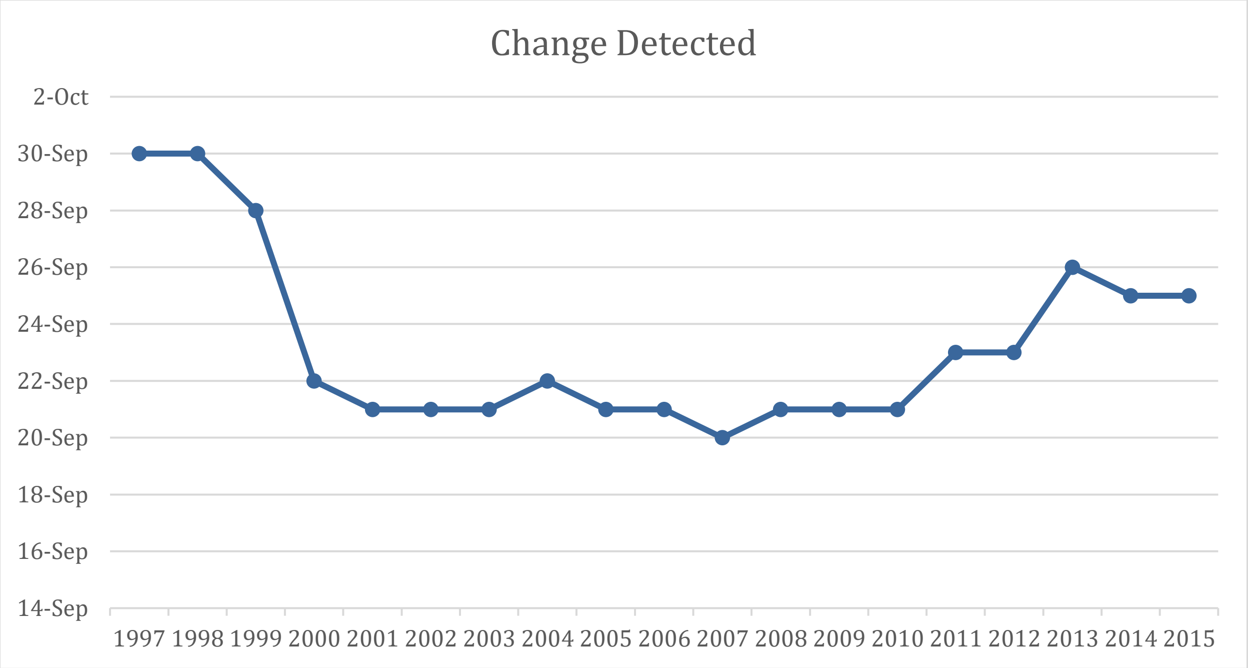
CUSUM for *xhat* output over 19 years and 123 days from Jul to Oct was performed with C = 1 times σ and T = 11 times σ. I had to pick a much higher threshold to detect consistent change because lower T values were detecting incorrect changes in short periods earlier. This analysis showed that summer had not ended sooner or later over 19 years. Summer ended at different times from mid Aug to mid Oct over the years with no indication that summer end moved later in the year.



CUSUM for *level* output was least sensitive and did not show much change over the years. Change was detected for only a handful of years for the same C and T values as xhat data (I did try other values too but result did not change much).

Day	Change
1997	7-Sep
1998	No Change
1999	No Change
2000	No Change
2001	30-Oct
2002	26-Oct
2003	No Change
2004	No Change
2005	29-Oct
2006	28-Oct
2007	No Change
2008	30-Oct
2009	No Change
2010	No Change
2011	No Change
2012	No Change
2013	No Change
2014	No Change
2015	No Change

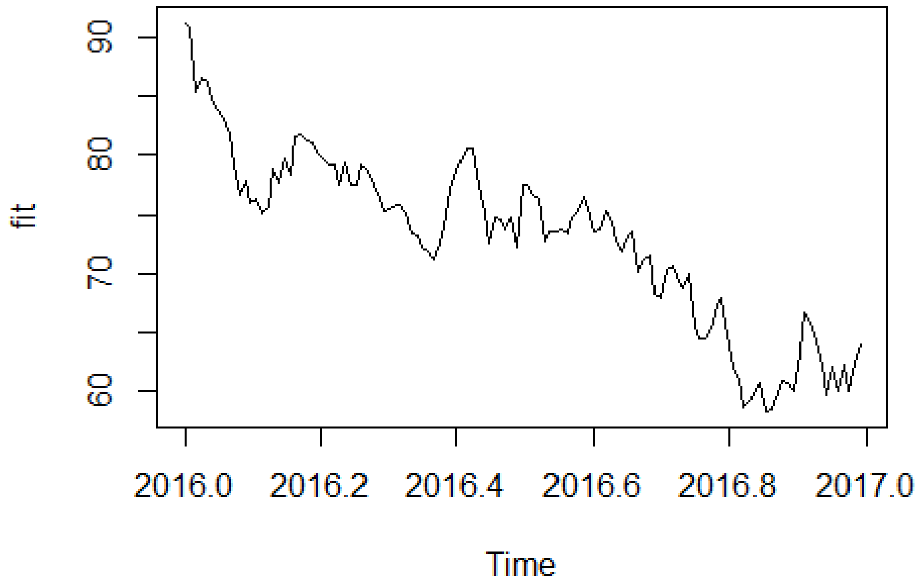
Lastly, CUSUM on *seasonality* data for C = 1.5 times σ and T = 5 times σ wasn't much different than previous analysis. If anything it showed that summer was gradually moving earlier until 2010/2011 but then ended a little later for the last few years. And the range in which summer ended is very narrow (20 or so days of Sept). Thus, I could not definitely say that summer ended later over the years.



More details are in the attached excel file.

I also used the Forecast library to predict the temperatures for the year 2016 using the tripple multiplicative holtwinters model. Intent was to see if the predictions for 2016 show any significant uptick in temperatures after Sept but the plot below did not show that. It showed temperature drops around end of Sept which was consistent with the other years.

```
#forecasting the whole 2016 using the multiplicative tripple exponential
smoothing model
model4_forecast = predict(model4, n.ahead = 123, prediction.interval = FALSE
)
plot(model4_forecast)
```



Before I closed this question, I wanted to check one more model. I implemented the ARIMA model below.

Condition of ARIMA model is that the data shouldbe stationary meaning mean, variance, and autocovariance should be time invariant. This assumption makes intuitive sense: Since ARIMA uses previous lags of series to model its behavior, modeling stable series with consistent properties involves less uncertainty. ADF Test checks whether data is stationary.

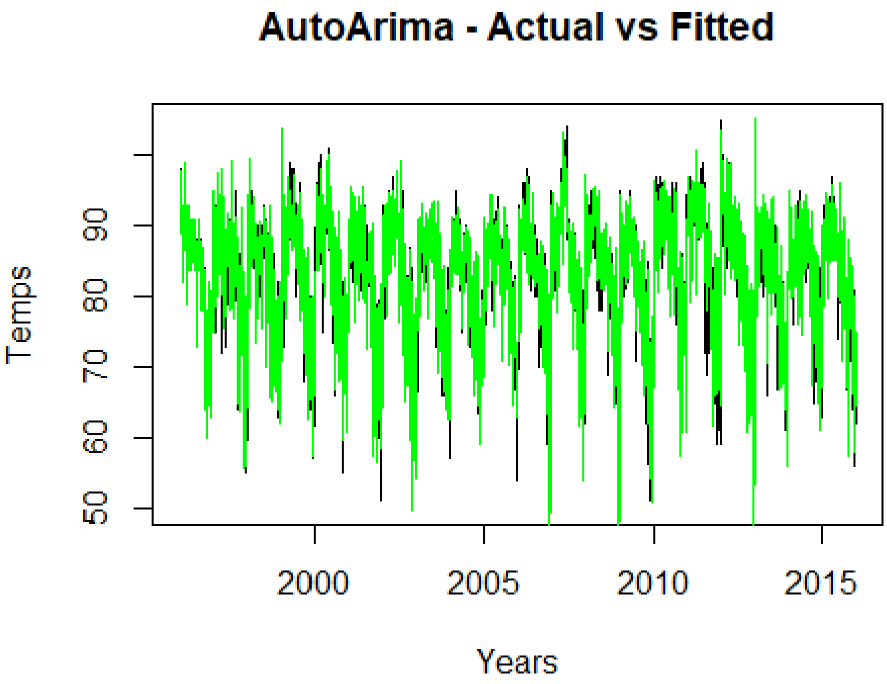
```
adf.test(ts_temps, alternative = "stationary")

##
##  Augmented Dickey-Fuller Test
##
## data:  ts_temps
## Dickey-Fuller = -7.4375, Lag order = 13, p-value = 0.01
## alternative hypothesis: stationary
```

A low P value indicated that NULL hypothesis that data was not stationary was rejected. I went ahead and ARIMA using AUto.Arima. Auto Arima returns a best model and picks best values for the order (p,d,q) and seasonal (P,D,Q) parts.

```
autoarima <- auto.arima(ts_temps)
autoarima

## Series: ts_temps
## ARIMA(1,0,1)(0,1,0)[123]
##
```

```
## Coefficients:
##          ar1     ma1
##       0.6423  0.0651
## s.e.  0.0229  0.0293
##
## sigma^2 estimated as 34.26:  log likelihood=-7444.75
## AIC=14895.5   AICc=14895.51   BIC=14912.77
```

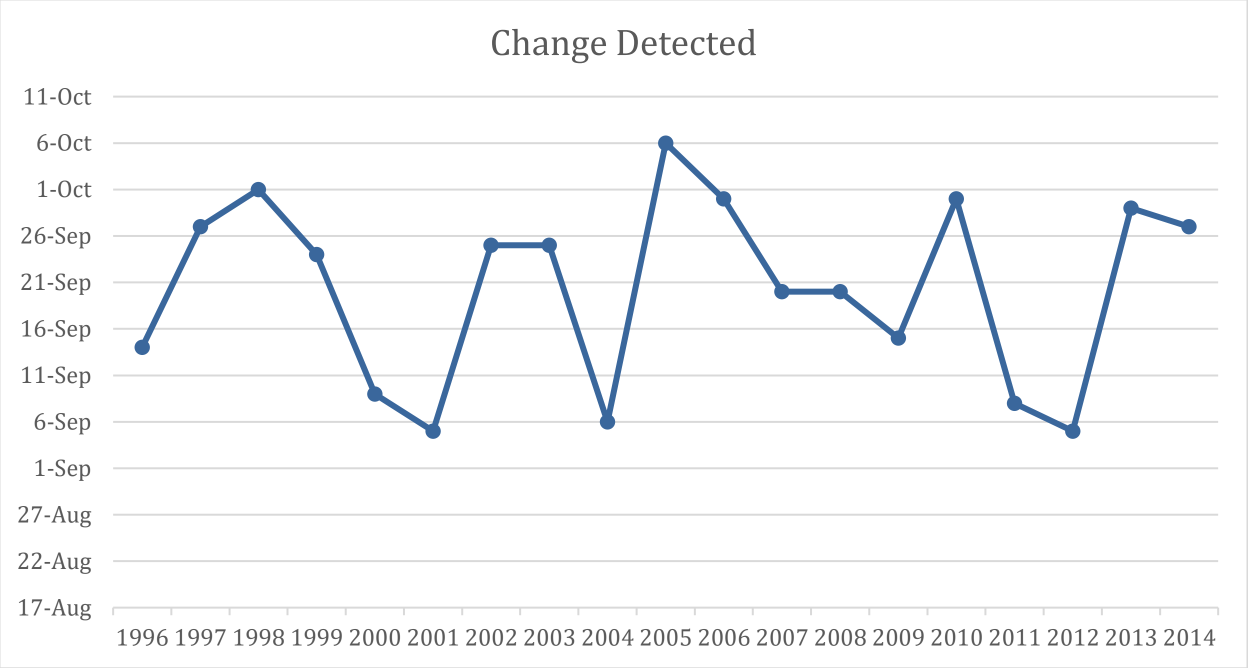
```
plot(autoarima$x,type="l",col="black", xlab="Years", ylab="Temps",
main="AutoArima - Actual vs Fitted")
lines(autoarima$fitted,col="green")
```



I exported the fitted values of ARIMA model to CSV and performed CUSUM on that data.

```
matrix_model5 <- matrix(autoarima$fitted, nrow=123)
write.csv(matrix_model5, file="7_1_arima.csv")
```

CUSUM on the ARIMA fitted values did not reveal anything different than tripple holtwinters. Summerended from early to end of Sep through out 20 years with no indication of summer ending later from this data as well.


Change Detected
11-Oct
6-Oct
1-Oct
26-Sep
21-Sep
16-Sep
11-Sep
6-Sep
1-Sep
27-Aug
22-Aug
17-Aug
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014