ISYE 6501 - Homework 4

2024-09-18

Homework 3

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 (the first smoothing parameter) to be closer to 0 or 1, and why?

I work for an ISP (Internet Service Provider) and need to predict signups from prospective customers. A signup signals the intent to become a paying customer. A prospective customer signs up via a one of our organic or paid channels and then either self installs or schedules one of our technicians to install internet service at their home. It is important to understand demand/signup volume for the following reasons:

-Anticipate the number of new paid customers we will get

-Ensuring we have the correct number of installer staff

-Order the correct amount of equipment (routers, network boxes, etc.)

-Staff up our call center correctly to handle the volume of inbound calls-

The data follows seasonal and market trends (i.e. higher signup volume in summer months and back to school). But every now and again there are irregularities in the data which can be due to higher deployment of sales staff on the field or on our call center in an effort to get more signups OR we extend our network to more territories and enable new prospects to sign up for service (inventory delivery). These events happen in spurts and it can be hard to predict the impact on signups. In order to provide an outlook to our management teams, we need to provide a smoothed, "situation normal", forecast.

In order to do this, we would need historical signup delivery by day, seasonality factors, inventory delivery schedule (to understand the magnitude of potential error), and market mix (as some markets deliver more customers than others due to competitive presence, size, etc.). As these expansion events are pretty random, I would expect the alpha to be closer to 1 as we can put more stock in previous data points to inform future dates. However, in months with higher inventory delivery and signup blitzes, I would expect an alpha closer to zero for recent data points.

Question 7.2

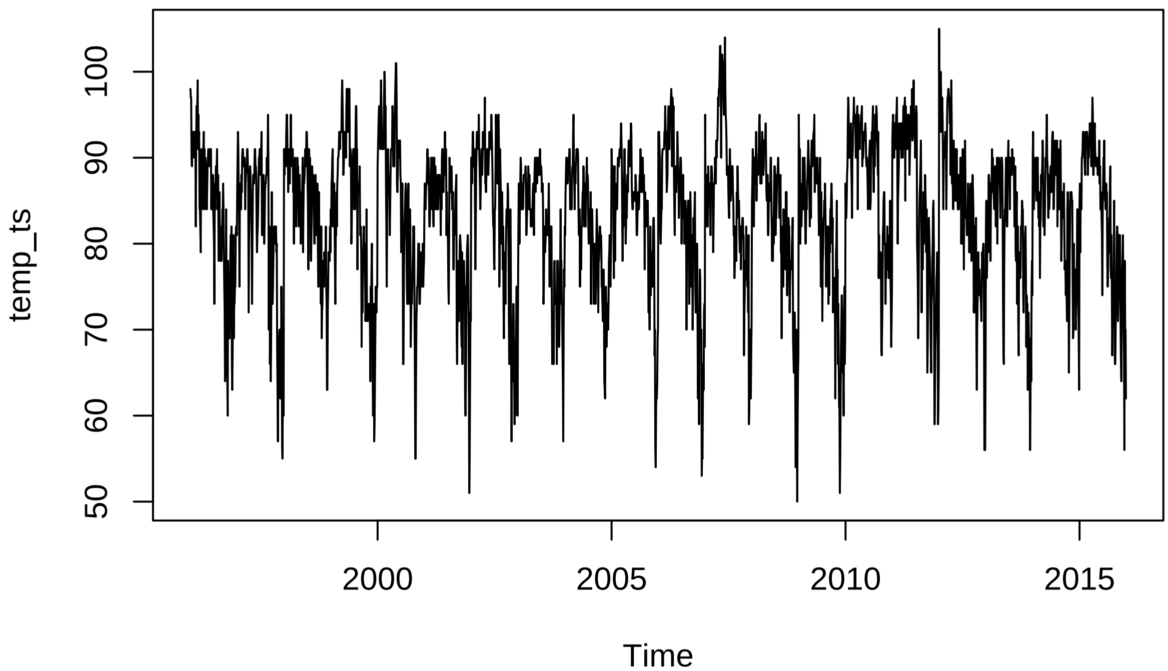
Using the 20 years of daily high temperature data for Atlanta (July through October) from Question 6.2 (file temps.txt), build and use an exponential smoothing model to help make a judgment of whether the unofficial end of summer has gotten later over the 20 years. (Part of the point of this assignment 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.)

I'm going to start by loading the necessary packages and data and then converting the dataframe into a time series. Then will visualize the timeseries

```r
library(tidyverse)
library(stats)
```

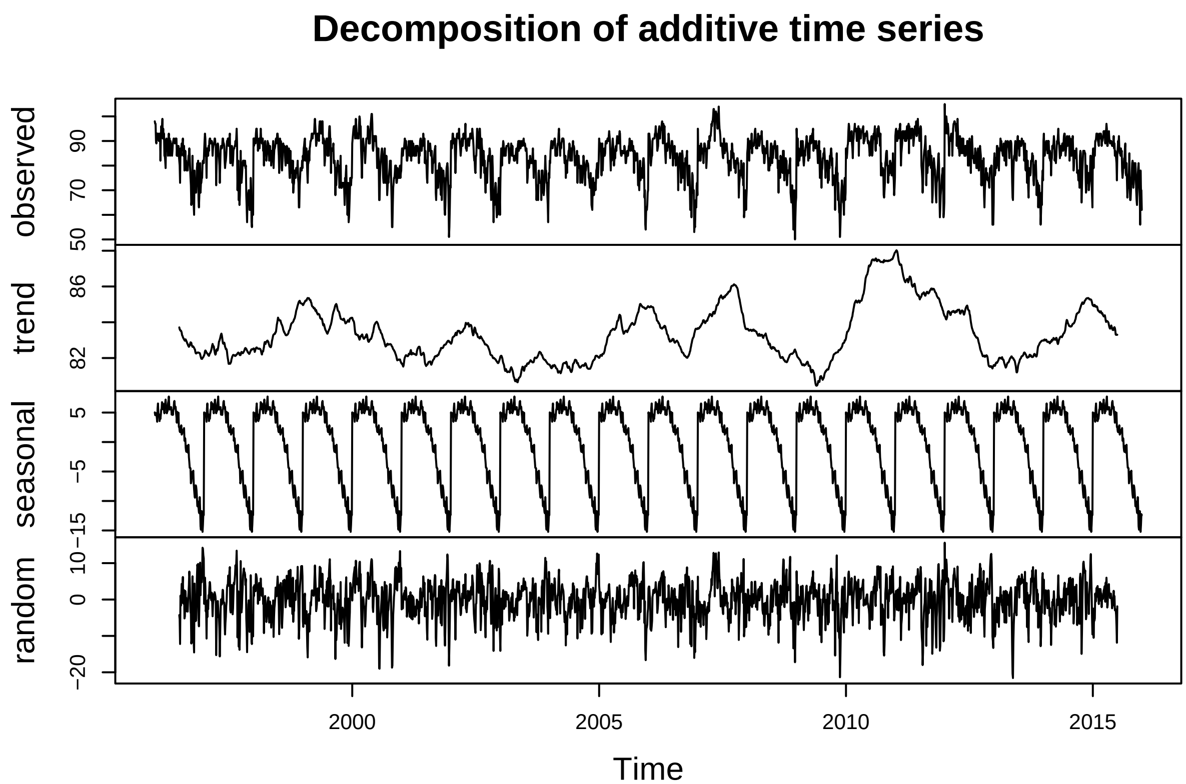
```r
data <- read.table("./temps.txt", header = TRUE, sep = "\t",
    check.names = FALSE, row.names = 1)
```

```r
temps <- data[, 1:20]
temps_vec <- as.vector(unlist(temps))
temp_ts <- ts(temps_vec, start = 1996, frequency = 123)
plot(temp_ts)
```



Next, I'm going to decompose the time series to see if there are any noticeable trends.

```r
plot(decompose(temp_ts))
```



Beyond the expected seasonality, we see the trend growing from 2010-2012. We don't see the trend steadily growing over time or seeing a changing seasonal factor. Seeing how random the data can be in a given period, I'm inclined to believe that the alpha should be closer to 0.

Next, I'm going to apply the HoltWinters function without any parameters to the entire data set and see what it provides. When looking at seasonality, I think the correct type to use here would be "additive" because the seasonal shift seems pretty consistent over time.

```
model <- HoltWinters(temp_ts, alpha = NULL, beta = FALSE, gamma = NULL,
    seasonal = "additive")
model$alpha
```

```
##     alpha
## 0.6610655
```

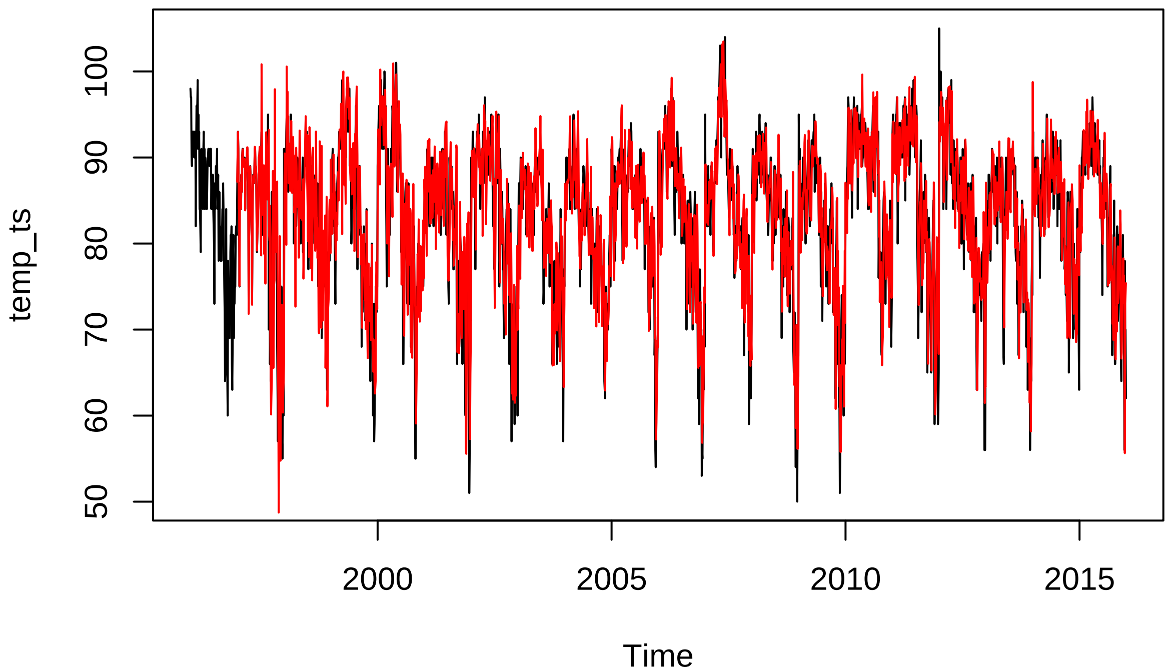
```
model$beta
```

```
## [1] FALSE
```

```
model$gamma
```

```
##    gamma
## 0.624816
```

```r
plot(temp_ts)
lines(model$fitted[, 1], col = "red")
```



When not specifying a parameter, the alpha is .66. This seems like it is over fitting the model and incorporating too much of the randomness. I'm going to do this again using .2 for alpha.

```r
model <- HoltWinters(temp_ts, alpha = 0.2, beta = FALSE, gamma = NULL,
    seasonal = "additive")
model$alpha
```

```
## [1] 0.2
```

```r
model$beta
```

```
## [1] FALSE
```

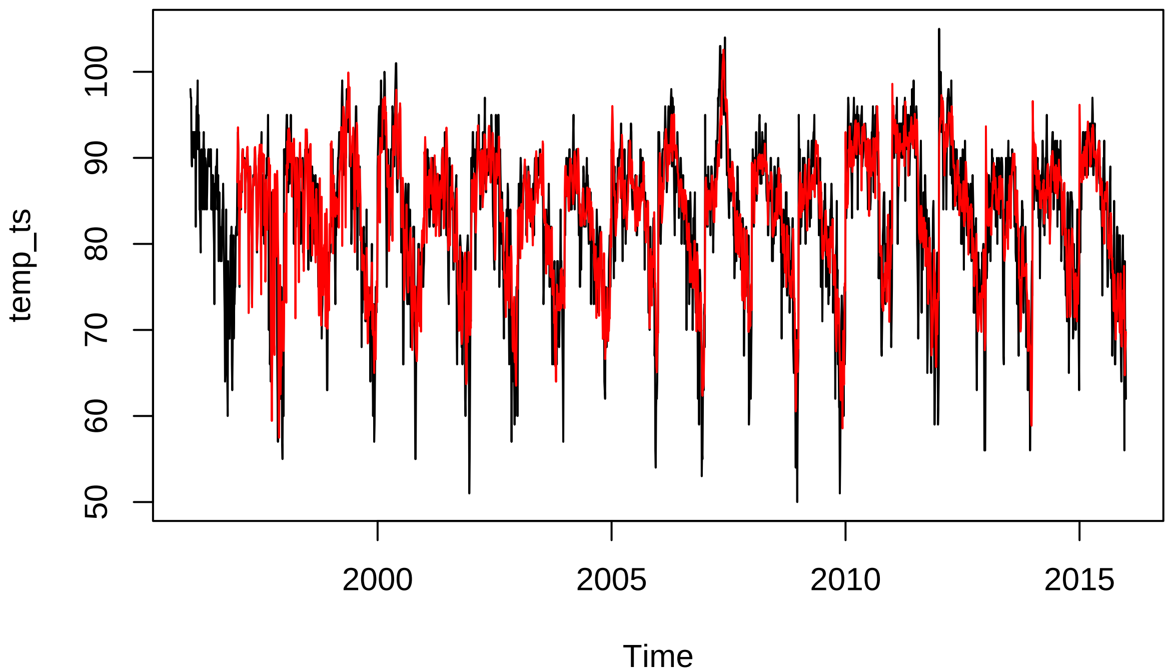
```r
model$gamma
```

```
## [1] 0.3206171
```

```r
model$seasonal
```

```
## [1] "additive"
```

```r
plot(temp_ts)
lines(model$fitted[, 1], col = "red")
```



The next step will be using the approach to loop through the original data frame and get smoothed results.

```r
results <- c()
for (i in 1:ncol(data)) {
    data_ts <- ts(unlist(data[, i], use.names = FALSE), start = 1,
        end = 5, frequency = 31)

    hw_model <- HoltWinters(data_ts, alpha = 0.2, beta = FALSE,
        gamma = FALSE, seasonal = "additive")
    results[[i]] = as.numeric(hw_model$fitted[, 1])

}
df <- as.data.frame(do.call(cbind, results))
colnames(df) <- colnames(data)
```

In order to see if summers have gotten longer over time, I was going to use the CUSUM approach from last week to loop through the data frame and get the point where the temperature crosses the threshold. Then I'll plot the minimum date by year in order to see if a trend exists.

Before running the model, I'm going to establish some constants. The C & T values will be set to the values I used to 5 & 10 respectively as it provided a reasonable estimate last week. The mu will be set to the average July temperatures.

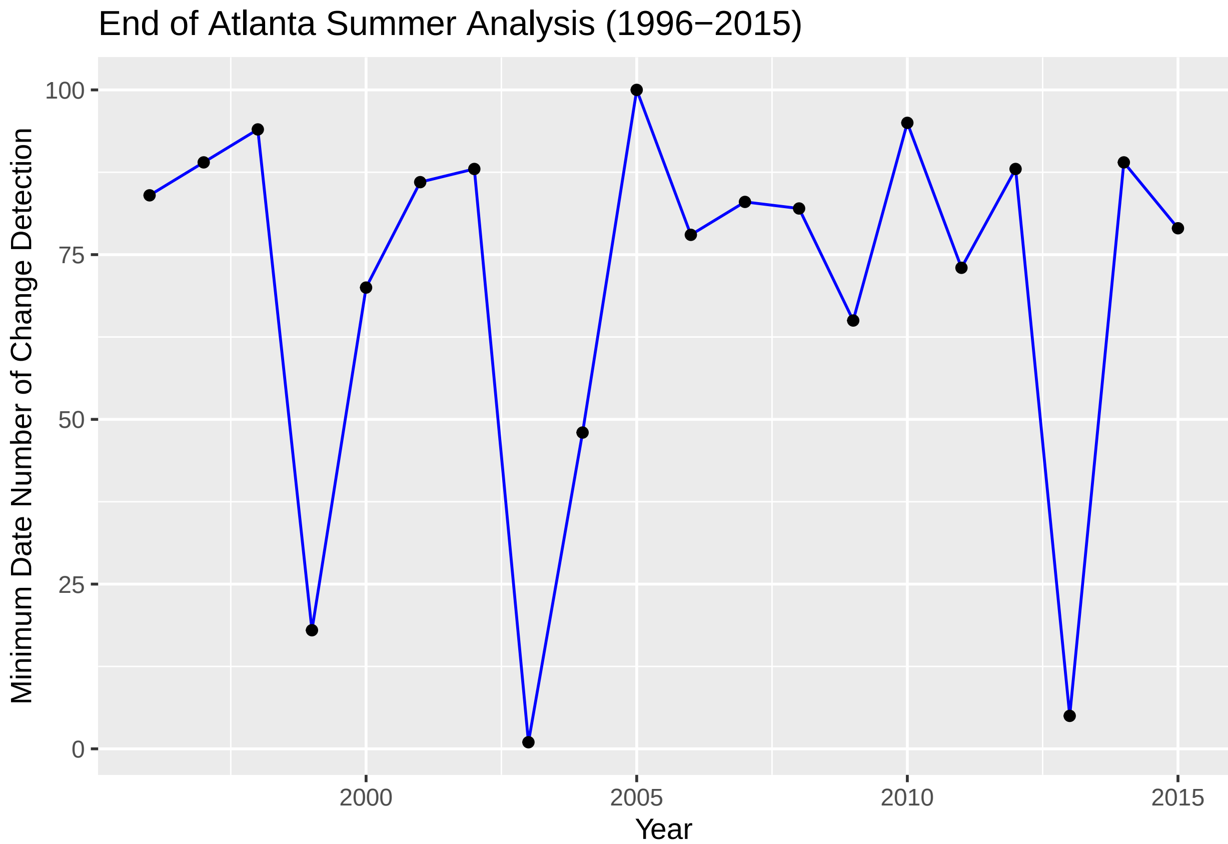
```r
data["num_date"] <- c(1:123)
mapping <- data.frame(date = rownames(data), data$num_date)
data_transposed <- data %>%
    gather(Year, value, -num_date) %>%
    left_join(mapping, by = c(num_date = "data.num_date"))
july_target <- data_transposed %>%
    filter(grepl("Jul", date) == TRUE) %>%
    summarise(avg = mean(value))
C <- 5
T <- 10
cusum_results <- data.frame(year = seq(1996, 2015), min_date = rep(0,
    20))
df["num_date"] <- c(1:124)
temp_df = data.frame(n_date = c(1:123), cusum = rep(0, 123))
```

I'm then going to loop through the data frame df and temporarily assign the CUSUM values of each year to the temp_df data frame. I'm then going to store the value of the minimum date where the CUSUM value crosses the threshold in the cusum_results data frame.

```r
for (i in 1:20) {
    temp_df = data.frame(n_date = c(1:123), cusum = rep(0, 123))
    for (j in 1:123) {
        if (j > 1) {
            temp_df$cusum[j] = max(0, temp_df$cusum[j - 1] +
                (july_target$avg - df[j, i] - C))
        } else {
            temp_df$cusum[j] = max(0, (july_target$avg - df[j,
                i] - C))
        }
    }
    # min_date = temp_df %>% filter(cusum>=T) %>%
    # summarise(min_date = min(num_date))
    temp_df %>%
        filter(cusum >= T) %>%
        summarise(min_date = min(n_date))
    cusum_results$min_date[i] = temp_df %>%
        filter(cusum >= T) %>%
        summarise(min_date = min(n_date))
}
cusum_results <- data.frame(year = cusum_results$year, min_date = unlist(cusum_results$min_date))
```

Finally, I'm going to visualize the data in order to determine if a trend exists.

```r
g <- cusum_results %>%
    ggplot(aes(year, min_date)) + geom_line(col = "blue") + geom_point() +
    labs(x = "Year", y = "Minimum Date Number of Change Detection",
        title = "End of Atlanta Summer Analysis (1996-2015)")
g
```



Looking at the graph, there is no indication, with the tests conducted, that summer ends later as time goes on.