Final Project: Simulation Functional Data Analysis

Functional Data Analysis With Application to United States Weather Data by Katherine S. King in 2014

Ruihan Jiang and Praveen Niranda Kumarasinghe Hetti Arachchige

MATH 590S, Spring 2023, Due Wednesday, May 3

Data Information

Similarities and differences in the use of data with the authors.

- Similarities:

Same Source:
National Oceanic and Atmospheric Administration (NOAA)'s National Climatic Data Center, from: https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/city/time-series/USW00024153/tavg/ann/12/1971-2023?base_prd=true&begbaseyear=1950&endbaseyear=2023

Same 16 Chosen Cities:
Burlington, LA, Portland, Miami, Salt Lake City, Nashville, NYC, San Antonio, Indianapolis, Minneapolis, Atlanta, Green Bay, Missoula, Fairbanks, Boston, San Francisco

- Differences:

Different Value:
Monthly average temperatures from 1971-01 to 2023-01 were used here, whereas daily temperature from 1950 to 2013 was applied in the paper.

5 Missing Data:
Missing data showed as '−99' in the raw data set, whereas paper used data without missing values, so we manually computed missing monthly average data from the website: https://www.extremeweatherwatch.com, then updated it.

Different Number of Time Points:
625 (= 12 months * 52 years + 1 month) were applied here, while there were $23,376$ data points used in the paper.

Different Period:
1 year (12 months) here, 365.25 days in the paper.

- Create Raw Data List

Referring to the data set growth from the lecture notes (file: fda.lec1.R), we store our data in a list named 'temp' with three sublists. Sublist one named 'date' is the year-month vector; sublist two is the temperature data matrix with 625 rows (each represents one time point) and 16 columns (each represents one city); and list three is the one-column Burlington city temperature data matrix.

```
### View parts of the raw data
temp$data[1:6,1:6]
```

```
##         Burlington Los.Angeles Portland Miami Salt.Lake.City Nashville
## 1971-01       12.7        53.9     40.5  67.6           32.8      35.3
## 1971-02       22.9        55.0     42.9  69.6           35.7      38.4
## 1971-03       27.0        55.1     43.8  69.6           40.8      44.3
## 1971-04       40.2        57.1     49.7  74.0           48.8      57.7
## 1971-05       57.4        59.7     56.9  79.4           57.7      63.0
## 1971-06       67.3        63.0     60.3  81.8           68.6      77.3
```

Transformation to Functional Data

The following steps were used and obtain figures matching those in the paper with overlapping time points, so it was reasonable to infer that the non-overlapping parts also made sense, and thus, the raw data was correctly transform to the functional data without the effects of seasonal variation.

- Step 1: Choose number of basis function ($K = 6$) to fit fourier series.

Although $min\{SSR\}$ was obtained when $K = 46$, there was no big difference between SSRs at $K = 46$ and $K = 6$ (see Figure 1 below) which means when adding more basis functions the sum doesn't substantially decrease, so it was not worth the cost of complicating the model, that was why $K = 6$ was chosen. Plots of seasonal fitting with fourier basis were periodic and didn't vary from year to year in its period.

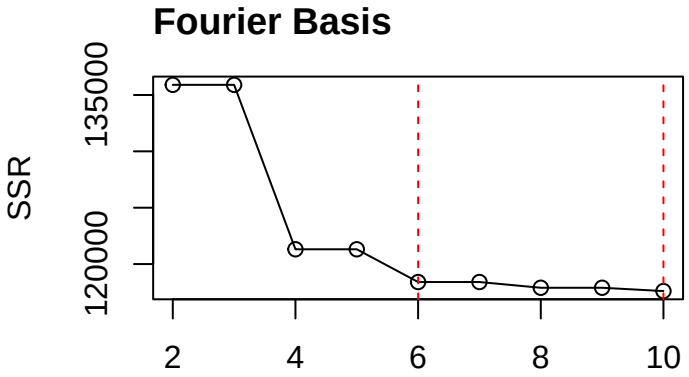

Figure 1

```
## [1] "done"
```

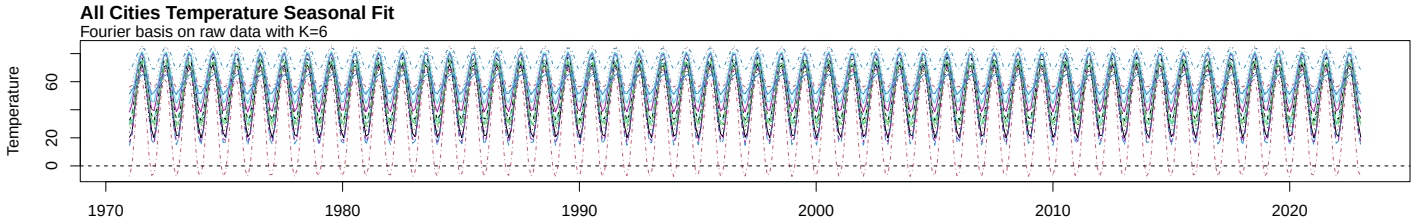

Figure 2

```
## [1] "done"
```

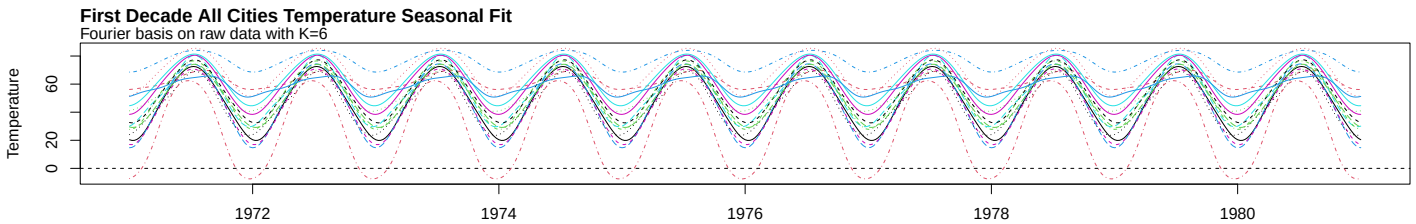

Figure 3

```
## [1] "done"
```

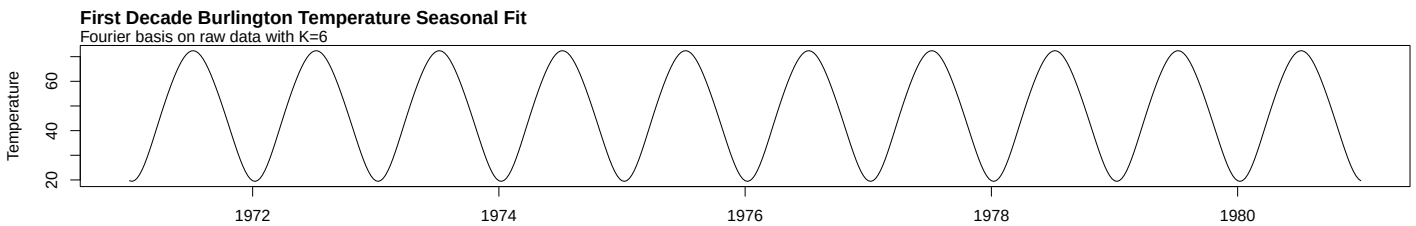

Figure 4

- Step 2: Choose $K = 109$ basis functions to fit the order 6 B-spline basis to the raw data.

Order 6 was chosen in order to estimate the first and the second derivatives well (smooth), the spline must have order four more than the derivative intended to be examined. Although $min\{SSR\}$ was obtained when $K = 150$, there was no big difference between SSRs at $K = 150$ and $K = 109$ (see Figure 5 below). To keep the model as simple as possible, we chose $K = 109$. This result was consistent with the result from the author since she also had the plot with the wired drop as what we got in Figure 5. Plots of seasonal fitting with B-spline basis were periodic too, but curves of B-spline basis plots slightly varied from year to year.

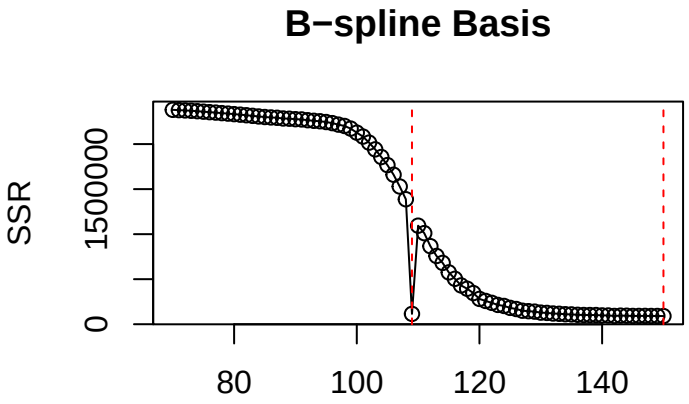

Figure 5

```
## [1] "done"
```

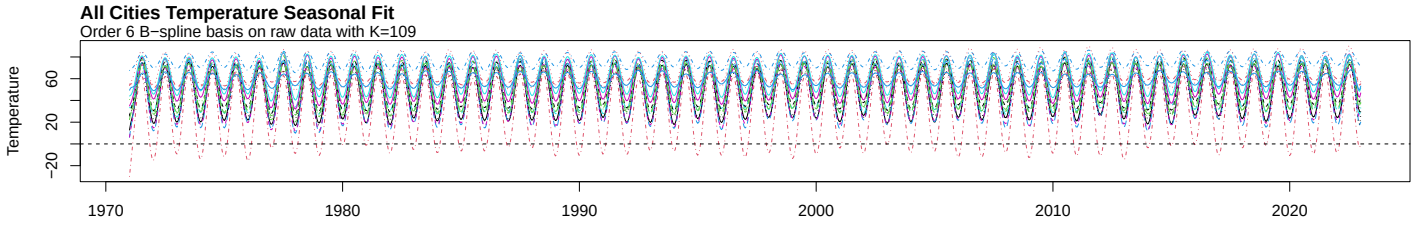

Figure 6

```
## [1] "done"
```

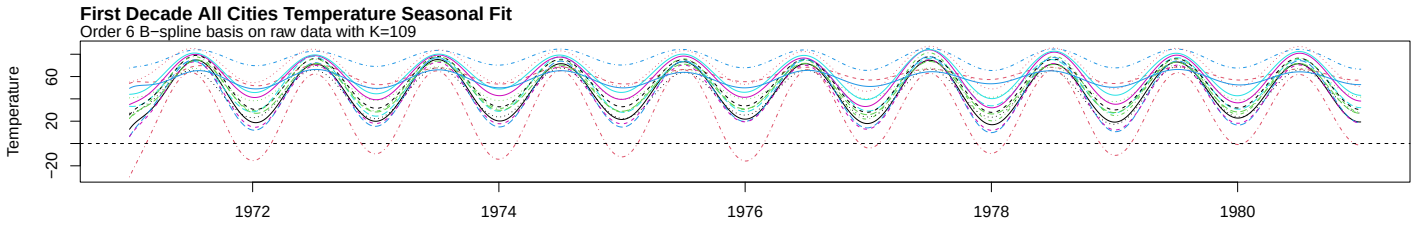

Figure 7

```
## [1] "done"
```

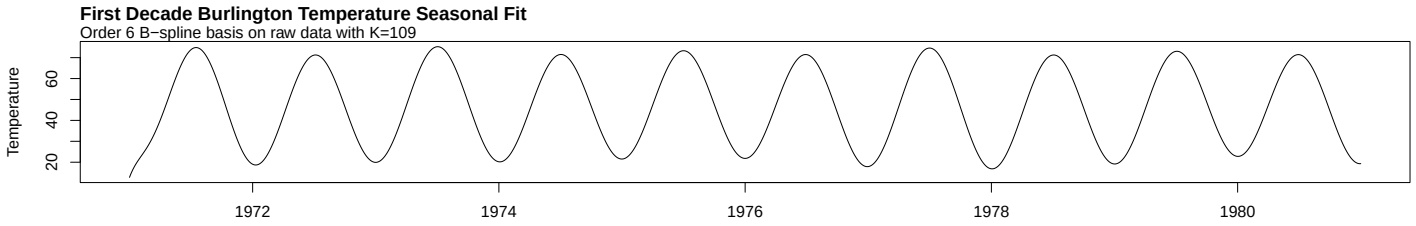

Figure 8

- Step 3: Subtract off the season component (Fourier series) from the previous spline using order 6 and $K = 10$ basis functions instead of $K = 109$ before penalty to "de-noise".

$K = 10$ was chosen in order to retain the shape of the curve, yet ensure the function did not exhibit excessive variation. The image now clearly shows us a trend of increasing temperature.

```
## [1] "done"
```

```
## [1] "done"
```

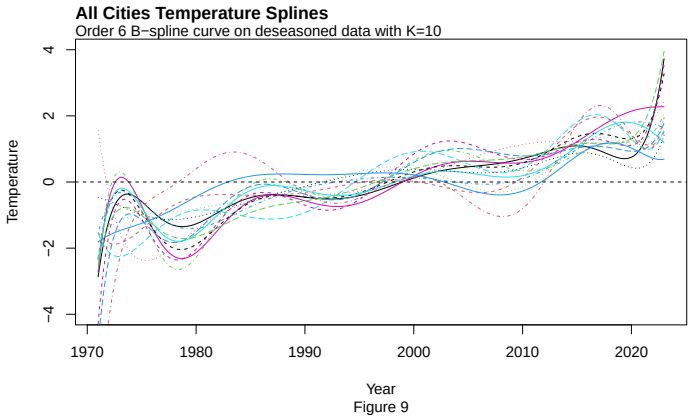

Figure 9

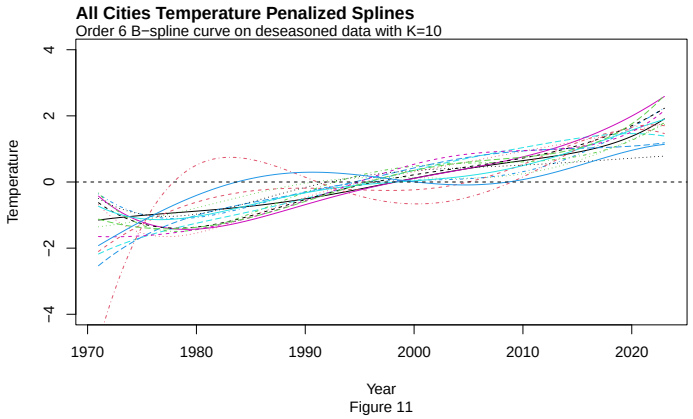

Figure 11

- Step 4: Use generalized cross validation criteria (GCV) to penalty the de-seasoned functional data.

The smooth parameter $\lambda \approx 10^6$ was chosen based on the lowest GCV score so that the residuals from the curves without seasonal variation was minimized as much as possible but still allowed for a smooth curve. Splines of all the cities with seasonal variation removed penalized against the 4th derivative using λ and $K = 10$ basis functions. The main smoothing occurred by decreasing the number of knots used so Figure 9 and Figure 10 were a little bit different, Splines after penalty (Figure 9) was more smooth than those in Figure 8. We can observe that there is an overall increasing trend in all curves.

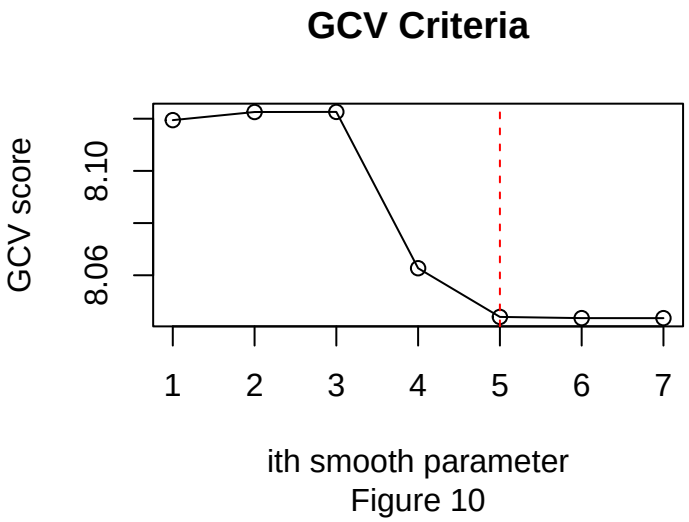

Figure 10

2 Units Standard Deviation Bounds (95% Confidence Limit) for Mean and Mean Derivative

Next we can look at the average of all cities together.

The standard deviation of the mean was doubled and added to both sides of the mean function to represent the mean with standard deviation limits above and below, shown in Figure 12 and Figure 13. Using a more precise calculation, the standard error covariance matrix was calculated and used to create a 95% confidence limit for the different cities in the study.

There is indeed an upward trend in the average temperature of all cities(Figure 12). The highest bound for average temperature in 1970 is significantly smaller than the lower bound for that in 2023. So the average temperature has gone up by around 2 degrees.

Furthermore the mean function is not a straight line so this pattern of upward momentum is not fully linear.

When observing the rate of change of the average temperature (Figure 13) we can see that it undergoes some dynamics during the first and the last decades. Furthermore there the rate of change, although it's small, it's always positive after the late 1970s. This further confirms the rise of average temperature over time. Moreover, between 1980 and 2000 even the lower bound for rate of change is positive meaning temperatures were in fact increasing at a significant rate.

We may look at temperature of individual cities for similar patterns (Figure 14 and 15).

```
## [1] "done"
```

```
## [1] "done"
```

```
## [1] "done"
```

```
## [1] "done"
```

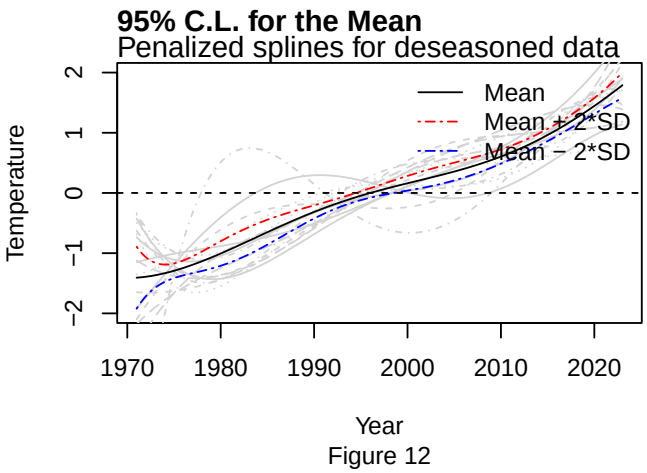

Figure 12

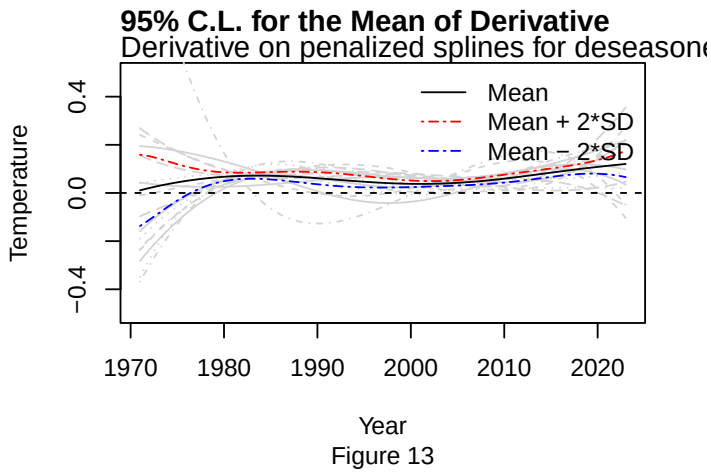

Figure 13

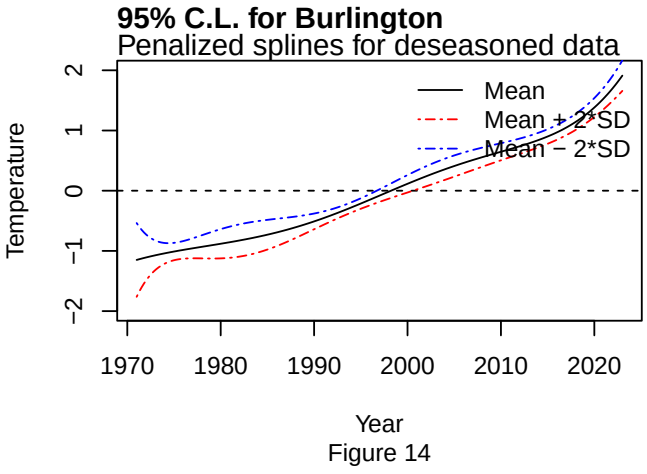

Figure 14

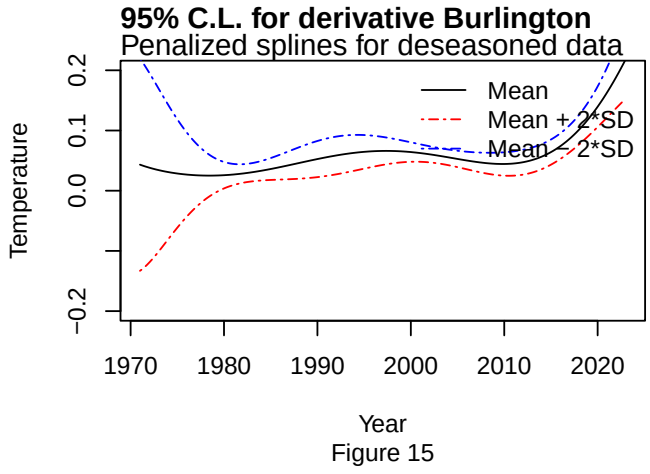

Figure 15

Permutation t-test

To see if there was a difference in the trends of the temperatures in various locations, a permutation test was employed to compare two sets of cities. Cities on the East Coast and cities on the West Coast were the groupings of interest. Cities on the East Coast are Burlington (VT), Miami (FL), New York City (NY), and Boston (MA). Cities on the West Coast are Los Angeles (CA), Portland (OR), Fairbanks (AK), and San Francisco (CA). Functional data plotted results showed in Figure 16 and Figure 17. It is obvious that both the temperature trend of two coast kept increasing from January 1971 to January 2023, although temperature trends of some western coastal cities had slight fluctuations.

```
## [1] "done"
```

```
## [1] "done"
```

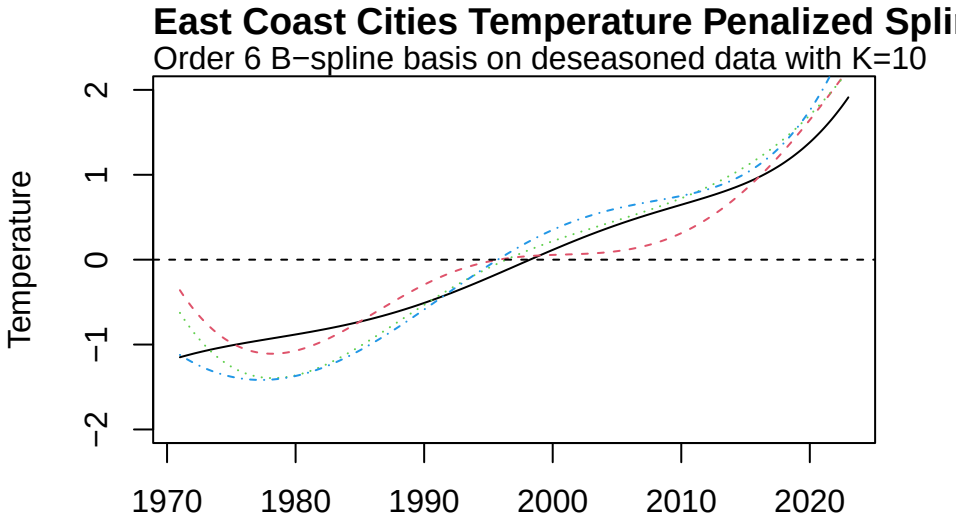

Figure 16

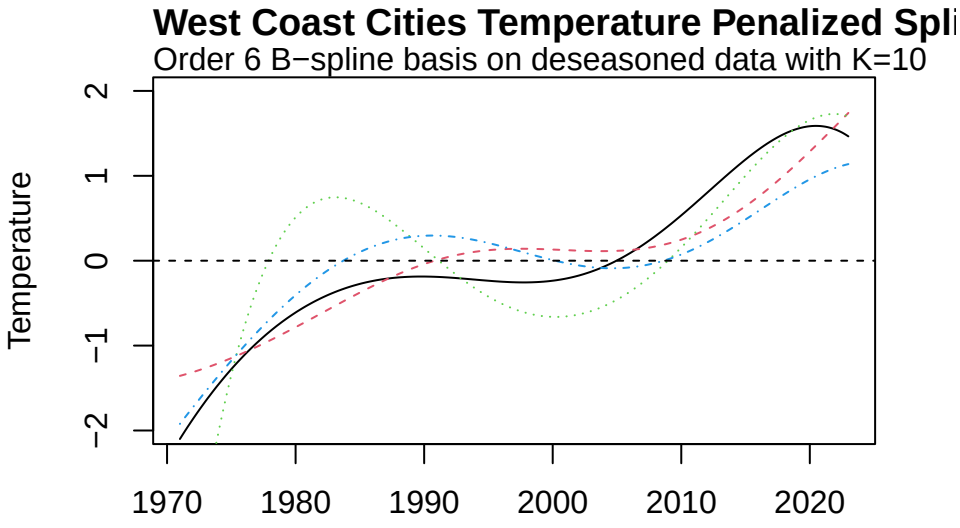

Figure 17

We rejected the null hypothesis (that two samples are the same) when observed statistics exceeded the point-wise critical value; otherwise, we did not succeed in doing so. Figure 18 demonstrated that two groups of cities occasionally had identical temperature trends, but occasionally did not. In this instance, our results were highly consistent with the author's findings, leading us to the conclusion that when combining the two coasts to form a comprehensive picture of the US, it may be appropriate to treat them separately rather than as a single unit, because they had overall similar increasing trend but couldn't be consider as the same.

```
## NULL
```

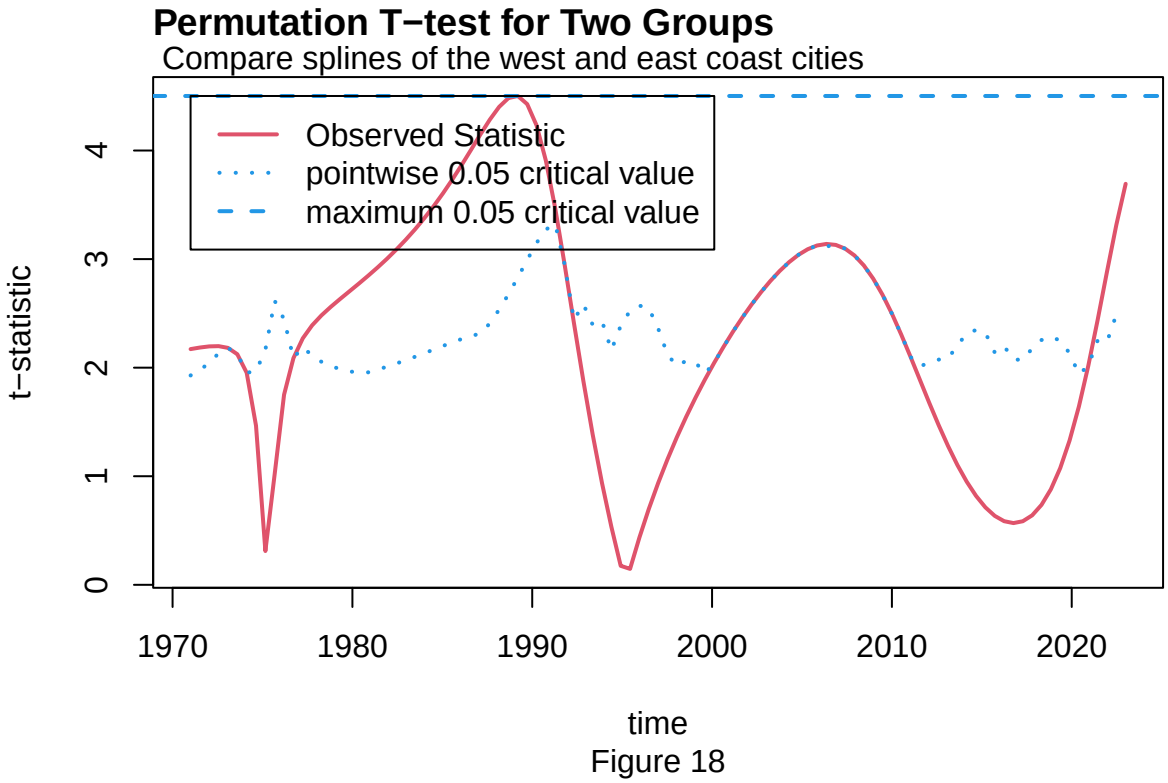

Figure 18

Summary

This study's main goal is to see if functional data analysis (FDA) techniques can be used to identify temperature variations in American cities during the past 60 years. The most difficult part of this simulation job was to remove seasonal effects and choose how to smoothing our data. Here we first fitted the fourier series to raw data with 6 basis functions and order 6 B-spline function with 109 basis function, then we took the difference to remove seasonal effect; finally we penalized the splines with fourth derivative and a smooth parameter. Then, we created 95% confidence limit for both mean and mean of derivative of de-seasoned functional data which also verified the increasing trend. In the end, we did the permutation t-test for two groups of cities, the findings indicate that temperatures have risen significantly in American cities over the past six decades which was not affected by the city's location on the east or west coast, and that the mean annual growth rate, which could represent overall annual temperature growth in the United States has been persistently above zero since the 1970s.