2014

Functional Data Analysis With Application to United States Weather Data

Katherine S. King
University of Vermont

Functional Data Analysis With Application to United States Weather Data

Katherine King

May 13, 2014

1 Abstract

This thesis explores the use of functional data analytic methods to examine climate change in a select group of 16 cities in the United States. The purpose of the project was to explore methods of functional data analysis in the context of climate change. Data used in this study was collected from NOAA's National Climatic Data Center. Major cities from around the United States were selected provided they had 100 percent coverage for the span of years of interest, 1950 through 2013. Sixteen cities were found to have complete data for every day of the 64-year period. Spline functions were fit to the temperature time series after removing seasonal variation. Mean temperature curves and associated confidence limits were computed. The results show a significant rise of temperature in U.S. cities within the last few decades and that the rate of increase has consistently stayed above zero since the 1970s.

2 Introduction

The Intergovernmental Panel on Climate Change defines climate change as "a change in the state of the climate that can be identified (e.g. using statistical tests) by changes in the mean and/or the variability of its properties, and that persists for an extended period, typically decades or longer[8]." Climate change is one of the most pressing concerns the world faces as a global community. Numerous studies are published each year to describe the state of the climate on global, national, and even local scales. According to the United States Global Change Research Program, the United States average temperature has risen more than two degrees Fahrenheit (approximately one degree Celsius) in the last 50 years [4]. This is more varied and extreme than overall global warming because oceans absorb more heat than air, slowing down the increase of surface temperature throughout the world[11]. Researchers find the shift of temperature detrimental to all areas of concern in the world. There is significant extinction of species due to the loss of habitat and resources from these new weather patterns and more are sure to follow[10]. The increasing temperatures and shifting climate patterns are a global crisis, which soon no one will be able to deny.

The overall aim of this study is to determine if changes in temperature of United States cities in the last 60 years is detectable using methods of functional data analysis (FDA). The methods of this paper transform discrete data into functions. In particular, the study will examine temperature trends over time after removing the effects of seasonal variation. It will compare trends using both an analysis on the function itself and its derivative. This will allow for a further interpretation of the speed and timing of temperature changes. The statistical methodology of functional data analysis is explained in the background section to provide the statistical model, assumptions, and theory behind the methods used. The methodology describes the transformation of the data into the functional form and the specific decisions made based on the data. Finally, the results section includes the statistical analysis and the functional plots of the data and concludes with interpretation and a discussion about future implications.

3 Background on FDA

In Functional data analysis (FDA) the data object to be analyzed is a function rather than a single data point. Using functions as data objects necessitates entirely new methods of analysis. There are many potential applications for FDA, including the analysis of longitudinal data. An advantage of FDA is it requires fewer assumptions than other methods of analyzing data over time. Many statistical methods for ”profile analysis” have been developed, but these methods either require assumptions on the specific functional form of the data over time, rely on repeatedly analyzing data at separate time points or require data to be measured at the same time point for each unit in the study. FDA also provides a richer set of analyses than just comparing means. With functions, it is possible to see and estimate trends, rates of change, and acceleration.

FDA starts with transforming data points to continuous functions, which is typically accomplished using spline functions or Fourier series. These functions are represented as

$$y_{ij} \quad = \quad x_i(t_{ij}) + \epsilon_{ij}$$

where t is time (or another continuous variable), y_{ij} is the jth observation of the ith sample function, and x_i is a smooth function. This smooth function is defined as

$$x_i(t_{ij}) \quad = \quad \sum_{k=1}^{K} \phi_k(t_{ij}) c_{ik}.$$

The $\phi_k(t)$ are basis functions, which are the building blocks for constructing functions. The c_{ik} are the coefficients associated with these K basis functions for the ith function and are estimated by the data.

Two types of basis functions are employed in this analysis, Fourier basis functions and B-spline basis functions. Fourier basis functions are limited in their usage since they only represent periodic data. Fourier basis functions are useful for examining annual trends with seasonal variation. The set of basis functions for Fourier series includes one constant function and then pairs of sine and cosine functions to capture the variation in phase (this implies the number of bases must always be odd).

$$\begin{aligned}
\phi_1(t) &= 1 \\
\phi_2(t) &= \sin(t\omega) \\
\phi_3(t) &= \cos(t\omega) \\
\phi_k(t) &= \sin(\frac{k}{2}t\omega) \\
\phi_{k+1}(t) &= \cos(\frac{k}{2}t\omega)
\end{aligned}$$

where ϕ_k is the kth basis function and $\omega = 2\pi/T$ where T is the period of the function. When data points are equally spaced, the functions have advantageous computational properties.

Splines, on the other hand, are piecewise polynomial functions defined over intervals such that they are continuous at interval endpoints, referred to as *knots*. For splines, the number of basis functions is

$$the\ number\ of\ knots + order\ of\ the\ spline.$$

This means that there will usually be many more basis functions to represent the data than there would be if just using Fourier series. Cubic splines, which only involve basis functions of at most degree three, are common in many analyses because they arise through constrained optimization problems, which provide theoretical justification for their use. Cubic splines are continuous and differentiable. However, in order to estimate derivatives well, the spline must have order four more than the derivative intended to be examined. This ensures derivatives are smooth. Using order six splines guarantees that the first and the second derivatives can be estimated well [5].

Determining the number of basis functions is a delicate process. Several factors must be considered and in the end the researcher may need to make somewhat subjective decisions. One technique common to help determine the number of basis functions is comparable to testing the variability of least squares regression. This is done by minimizing the square of the residuals of the predicted x_i and the observed y_{ij}. The residuals should be penalized by the number of basis functions (in the case of splines, the number of basis functions is equal to the number of knots plus the order of the spline) so as to give more information in terms of variability. The residual calculation is

$$s_i^2 \;=\; \frac{1}{n_i - K}\sum_{j}^{n_i}[y_{ij}(t) - \hat{x}_i(t_{ij})]^2$$

where n_i is the number of observations for the ith curve and K is the number of basis functions used in the analysis. Adding basis functions will decrease the sum of squared error but at the expense of complicating the model. An appropriate number of basis functions can be determined by plotting $\sum s_i^2$ versus K and noting where the plot flattens, an indication that adding basis functions beyond this point is not improving model fit.

Fitting data using either splines or Fourier series results in smoothing of data. Great care must be taken not to over-smooth, thereby missing important features of the true underlying curve, nor to under-smooth, thereby estimating noise(random error). More knots means less smoothing.

Another way to determine the amount of smoothing is to choose many knots and add a roughness penalty λ to the integrated squared second derivative of the estimating function:

$$\text{PenSSE} \quad = \quad \sum_{j=1}^{n_i}(y_{ij} - x_i(t_{ij}))^2 + \lambda \int [D^2 x(t)]^2 \, \text{dt}.$$

The first term is the sum of square error for residuals where y_{ij} is fit by the curves of x_i, which were defined before as basis functions with coefficients specific to the data. Minimizing *only* the sum of squares is a least squares problem and can result in oversmoothing if there are many knots. Therefore a roughness penalty λ is added to the equation. The roughness of a function is commonly measured by the total curvature of the function, given by the integrated squared second derivative. If derivatives of $x(t)$ are of interest, increasing the degree of the derivative of the integrated penalty would ensure the functional derivative was smooth. In many cases, the fourth derivative is penalized to ensure acceleration of the function is smooth. For Fourier series, testing the curvature does not necessarily fit the purpose so instead the use of harmonic acceleration operator is suitable (although this paper does not go in depth with this topic). Determining the roughness penalty is not an exact science and should be evaluated in more than one way to make sure results are consistent. Generalized cross-validation is typically used to choose λ [5]. Define

$$\text{GCV}(\lambda) \quad = \quad \left(\frac{n}{n - df(\lambda)}\right)\left(\frac{SSE}{n - df(\lambda)}\right),$$

where $df(\lambda)$ measures the effectiev number of parameteres used to estimate $x_(t)$. [6]. Typically, a range of values around the value of λ minimizing

GCV(λ) give similar results and choosing an exact value adds some subjectivity to the analysis.

The last main topic of transforming the data before analysis is aligning the features of the curves, called registration. Data over time collected on different units can exhibit differences of amplitude and phase. Averaging over functions that are not aligned can dilute functional features. To align functions, time warping is used to slow down and speed up the phase of the curves. This can be done with landmark registration, which requires finding a feature of interest to align curves. An example would be to find the second derivative and then align every curve so they had zero acceleration at the same time. This is not too demanding if the functions are short and they are strictly increasing. But for larger data sets with periodic functions over a long stretch of time, the effort can be much more difficult.

Processing the data is the central backbone of functional data analysis and deserves a great deal of thought. When it comes time for analysis, the model assumptions must be verified and the data needs to be processed in a way that does not misrepresent the underlying true function. Descriptive statistics such as mean, standard deviations, and even confidence intervals can be analyzed once the data is properly fit. The mean and variance functions are:

$$\begin{aligned} \bar{x}(t) &= \frac{1}{N}\sum_{i=1}^{N} x_i(t) \\ s^2(t) &= \frac{1}{N-1}\sum_{i=1}^{N}[x_i(t) - \bar{x}(t)]^2 \end{aligned}$$

where N is the number of sample functions.

The standard deviation formula is just extended from other common statistical analysis. In order to get confidence limits for the different functions Σ_e, the covariance matrix of the residuals is estimated. Methods for estimation of Σ_e are described in Ramsey and Silverman [6]. Means are calculated from every sample function for every time step and then the coefficents of the basis functions were used. This allowed for a more accurate estimation of the standard error because instead of looking at discrete points it used the calculated coefficients. This error matrix is assumed to be diagonal so only the residuals across each curve for each observation need to be calculated. This helps with the estimation of the standard error so that a confidence for each curve, including the mean curve, can have a 95% point-wise confidence limit.

Derivatives can be important functional statistics as well since they measure rate of change. This is in fact one of the real benefits of using functional data analysis[3]. With discrete data, the ability to look at changes over time is very limited because analyses involve looking mostly at endpoints and describes little information inbetween unless strong assumptions on the parametric form of the model are made. In FDA, statistical software constructs derivatives from the original spline fits (this is based on the assumption that the smoothing parameter was chosen such that two more than the derivative of interest was used as a penalty).

Two-sample t-tests can be extended to FDA and are commonly empliyed as it is a common scenario to have data collected over time on individuals or units in two groups, and to then ask whether functional forms for the groups differ. Hall and Van Keilegom[2] go into great detail about the assumptions and the type of data necessary to compare two groups. They recommend equal smoothing parameters and subsamples of the groups at the same equally spaced locations for each curve. The null hypothesis is that the two samples come from identical populations. The fda package in R executes a two-sample t-test that outputs point-wise t statistics which determine where in the functions there exists a significant difference if any.

4 Methods

4.1 Collection

The data used in this study was collected from NOAA's National Climatic Data Center. Major cities from around the United States were selected provided they had 100 percent coverage for the span of years of interest, 1950 through 2013. The NOAA Data Center provided the minimum and maximum temperatures of different weather stations. In the end, sixteen cities were found to have complete data for every day of the 64-year period. (Burlington, VT, Los Angeles, CA, Portland, OR, Miami, FL, Salt Lake City, UT, Nashville, TN, New York City, NY, San Antonio, TX, Indianapolis, IN, Minneapolis, MN, Atlanta, GA, Green Bay, WI, Missoula, MT, Fairbanks, AK, Boston, MA, San Francisco, CA). The average of the minimum and maximum temperatures were used for analyses and in total there were 23376 time points (n_i were all equal). Most of the procedures that follow are from the package fda in R created by J. O. Ramsay et al [5].

4.2 Transforming the Data into Functions

To properly examine changes in temperature across cities, seasonal and random variation need to be removed. Seasonal variation can be removed by subtracting off the average median temperature of the cities at given times of the year (in this case everyday). There are many different ways to attempt this, for example finding an annual average using the first five years of data and subtracting that average from the full data set. Using functional data techniques, fitting a Fourier series to the full data seemed appropriate [7][6].The Fourier series used the entire data set to create a periodic function that repeats the same function every year (a period of 365.25 days since there are 16 leap years in the data) and used five basis functions. To determine the appropriate number of basis functions, the residuals of the predicted and the observed data points were minimized. Figure 1 plots the number of basis functions versus the residuals penalized by the number of basis functions (since raw residuals would naturally decrease). The biggest decline occurs when number of basis functions, k, is five and the sum of the residuals is 288. Although the plot does not go out further, the Fourier series with 101 basis functions has residuals that sum up to 283. This shows when adding more basis functions the sum doesnt substantially decrease and it is not worth the cost of complicating the model.

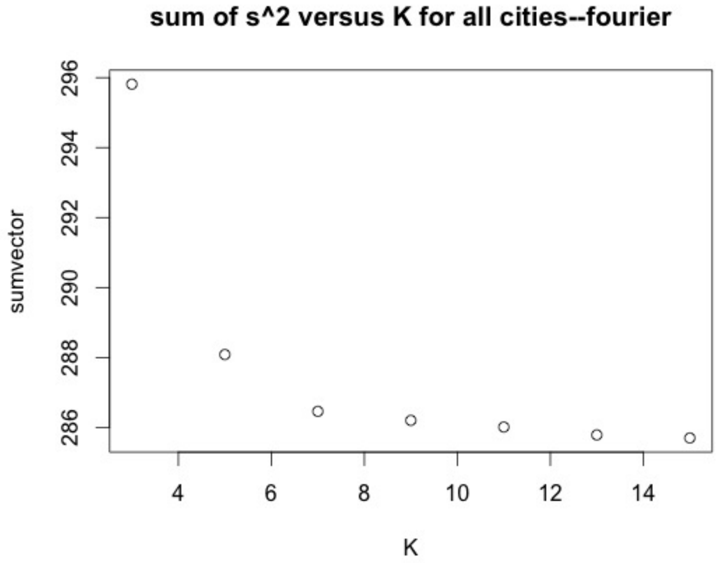


Figure 1: Sum of the Squared Residuals from the original data and the data predicted by a Fourier series. K is the number of basis functions required (the number of sine and cosine functions used) and sumvector is the total of the squared residuals added together.

Next, an order six spline was fit to the original data with 133 basis functions so the curves of the cities were smoother than the raw data. Figure 2 shows the sum of the residuals versus the number of basis functions similar to the calculation used for the Fourier series to determine the number of basis functions. With 133 basis functions, the sum decreases significantly and does not get back down to a sum of around 400 unless 20 more basis functions are added. This is a strange drop in the data since usually the sum should consistently decrease. One possible explanation for this drop at 133 basis functions is that 133 is twice the number of years plus the degree of the spline. It would make sense to have two knots each year, one representing the minimum value and the other the maximum value. So for future analysis 133 was chosen because it kept the model as simple as possible.

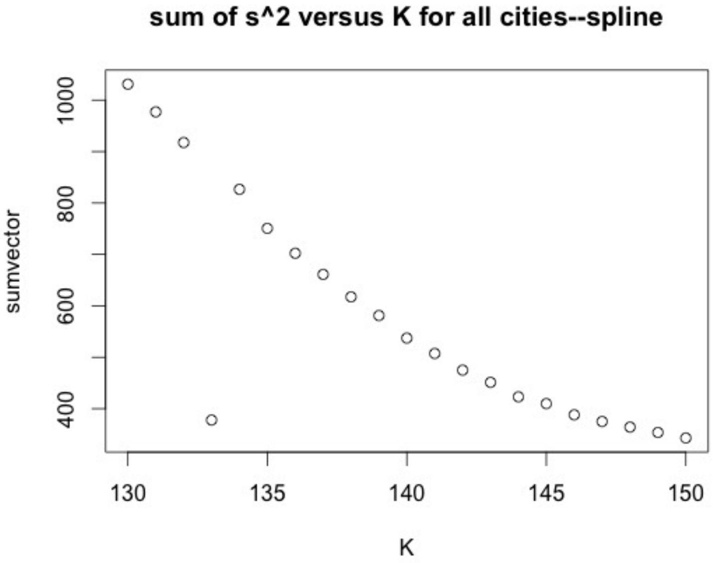


Figure 2: Sum of the Squared Residuals from the original data and the data predicted using an order 6 spline fitted to the data. K is the number of basis functions required (the number of knots plus the order of the splines used) and sumvector is the total of the squared residuals added together.

The seasonal fits (using Fourier basis) were subtracted off each time point of the spline fits to the raw data. This was done by evaluating both the Fourier series and spline functions at every time point over the 64 year period. The residuals of these two functions are now the main concern of analysis.

The residuals from the Fourier series and spline of the original data went through multiple subjective processes. This was taken into account when the residual data needed to be smoothed again. An order six spline with only ten basis functions was created using these residuals. It was important that the splines be order six so that subsequent analyses would have a smooth second derivative penalized against the fourth derivative. In this test, the number of basis functions was not determined from the plot of $\sum s_i^2$ vs k as described above. Instead, a minimal number of basis functions was used in order to "de-noise" the data before penalization to further smooth the data. Ten was chosen as the number of basis functions in order to retain the shape of the curve, yet ensure the function did not exhibit excessive variation. In addition to using a small number of basis functions, another smoothing technique was applied. The fourth derivative of the spline was penalized using a roughness penalty λ. Using generalized cross validation criteria, λ was chosen so that the residuals from the curves without seasonal variation was minimized as much as possible but still allowed for a smooth curve. Figure 3 shows the plot of the the $\log_{10}$GCV versus the $\log_{10}\lambda$ since the only interest is the scale of the number of digits. It is seen that increasing λ to even 10^{14} is not all that different from when λ is 10^{-2}. In order to follow through with GCV criteria, when $\lambda = 10^{10}$ the GCV value is minimized. This was the smoothing parameter used through the rest of the analysis[5].

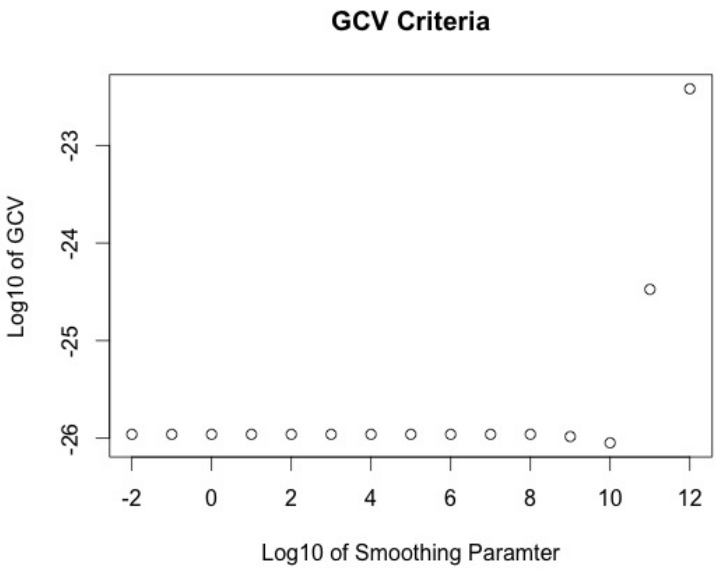


Figure 3: The sum of GCV over all cities versus the smoothing parameter λ penalized against the fourth derivative.

Before statistical analysis were performed, registration of the curves was considered. Registration is used to align the curves in phase and amplitude. This allows more information to be retained when calculating the mean of the curves because the functions match up more. This was found to be a challenging undertaking on this data set because there were no clear aligning characteristics of the curves. Therefore, the analysis was done without registering so the effects and timing of the curves may have been slightly muted.

4.3 Statistical Analysis

The functional mean and standard deviation of the 16 cities were computed and graphed. The mean and standard deviation were found using functions in the fda package in R built specifically for analyzing multiple functional curves. The standard deviation was calculated by summing the squared differences of the mean function and each individual curve. The standard error was calculated by estimating the covariance matrix of the error. This was estimated by smoothing mean square residuals of every station at every time point (every day) and then using the coefficients and the functions of the basis functions to determine the covariance matrix for the curve values, not just each data point. This standard error was then doubled and added to each curve so that each city at had a 95% confidence limit.

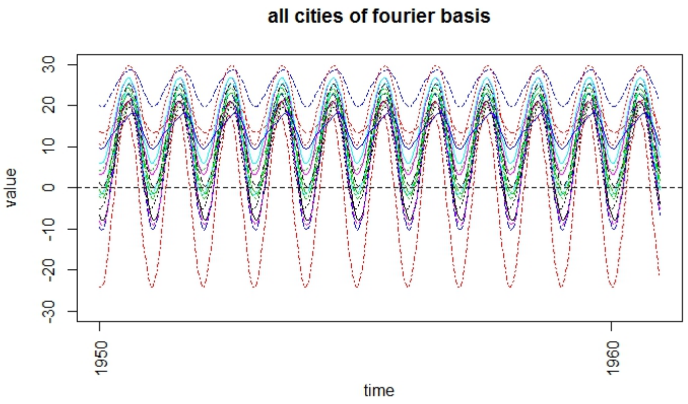
The next task was to examine the derivatives calculated using the function deriv.fd, which is also part of the fda package. The mean of the derivative functions were plotted. Using the standard deviation, based on the sum of the squared differences of the mean and actual derivatives, two standard deviations bounds were placed around the mean of the derivative.

Lastly, a two-sample *t*-test was utilized to evaluate the differences in the residuals between groups of cities defined by different geographic locations. The hypothesis tested was that there was no difference in temperature trends between the West Coast and East Coast of the United States. For the evaluation of differences between the coasts, only cities relatively close to the oceans were examined. East Coast cities were Burlington VT, Miami FL, New York City NY, and Boston MA while West Coast cities were Los Angeles CA, Portland OR, Fairbanks AK, and San Francisco CA. A t permutation test was calculated based on an R function in the fda package. This test looks compares the tests using both overall and pointwise

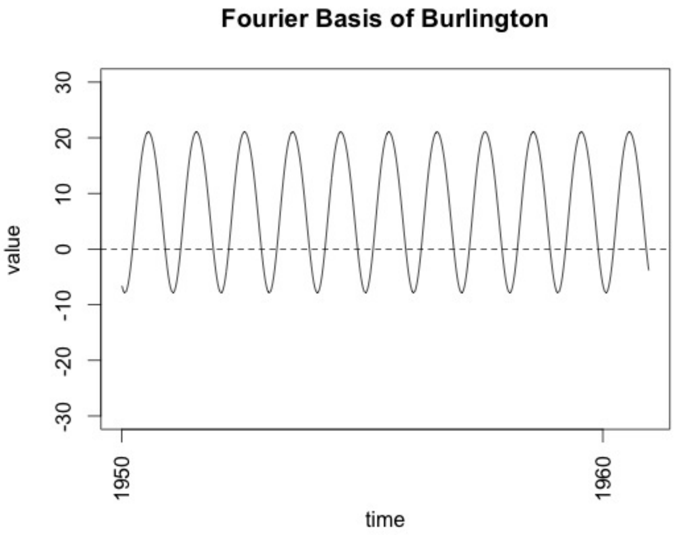
5 Results

As described above, seasonal variation was modeled using a Fourier series comprised of five basis functions. To provide adequate resolution, Figure 4 (a) shows the Fourier fits for all cities only for the first decade. Figure 4 (b) shows the Fourier series for only the city of Burlington VT.

Figure 4



(a) Fourier series of the original data using 5 basis functions. Each color represents one of the 16 cities. Only the first decade of the data is represented.

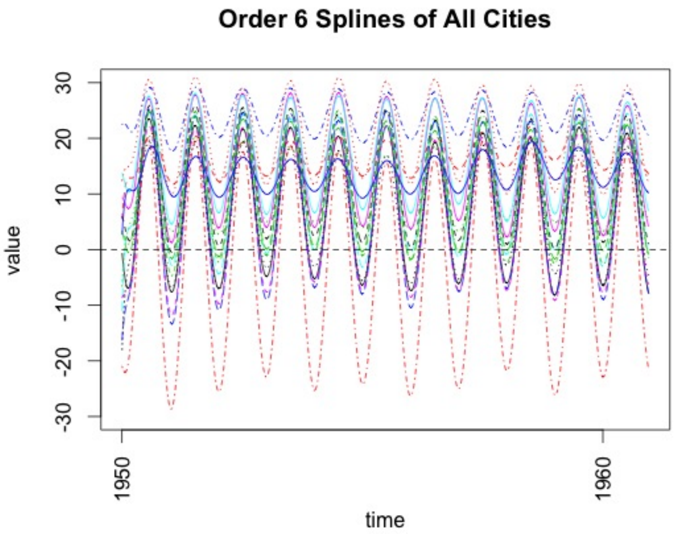


(b) Fourier Series of Burlington VTusing 5 basis functions. Only the first decade of the data is represented.

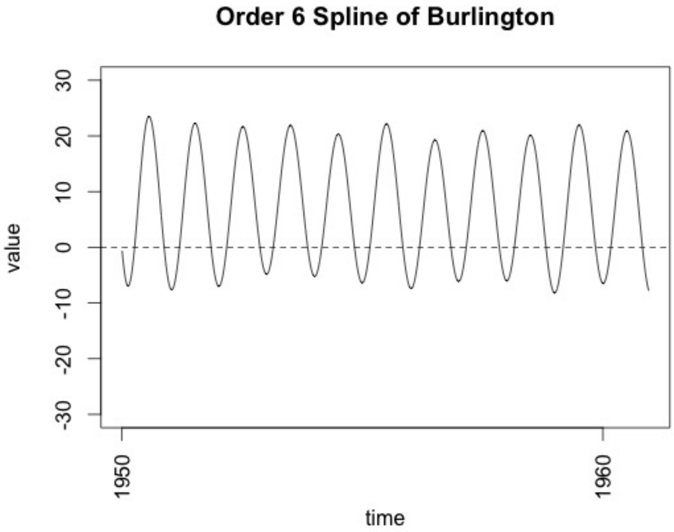
From the graph of only Burlington VT, it is easily observed that the function is purely periodic. It does not vary from year to year in its period, maximum value, minimum value, or even its phase. In essence this represents the average trend of temperature over the 64 years of interest in Burlington.

Deseaonalized data was obtained by subtracting the Fourier fits from a smooth of the original data. In order to remove additional noise, the spline was first fit to the original data and the season component (Fourier series) was subtracted off from this spline. Based on previous criteria, 133 basis functions were used with order 6 splines on the original data. The graph of the first decade of these splines are in Figure 5. The curves retain the seasonality but unlike the Fourier series the seasonality varies from year to year, reflecting changes in the timing of the onset of seasons and their severity. Instead of representing an average, the splines represent the actual trend of that specific year for each city. Figure 5 (b) of Burlington VT shows the minimum and maximum values of the function change based on each year of data. The period of some of the functions are shifted very slightly, though not apparent to the naked eye, which captures the difference in timing of the change of seasons from one year to the next.

Figure 5

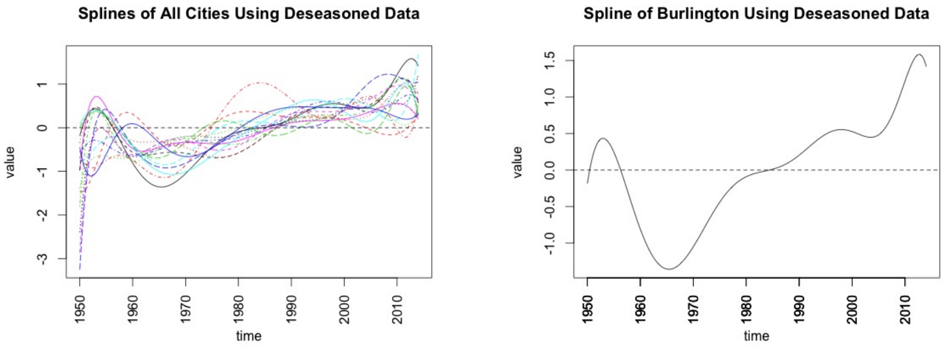


(a) Order 6 splines of the original data using 133 basis functions. Each color represents one of the 16 cities. Only the first decade of the data is represented.



(b) Order 6 splines of the Burlington VT using 133 basis functions. Only the first decade of the data is represented.

Using these two functional data objects, it was appropriate to find a new object based on their differences. For each city, a spline was fit to the differences of the Fourier series and the spline fit to the original data. However only 10 total basis functions were used in order to avoid having excess noise obscure the underlying trend in the temperature time series. Figure 6 (a) shows the splines of all the cities with the seasonal variation removed and (b) shows just the city of Burlington, VT. Figure 7 represents very similar functions except these were penalized using $\lambda = 10^{10}$ as the smoothing parameter. Figure 6 and Figure 7 are almost identical to the naked eye. This was expected because the main smoothing occured by decreasing the number of knots used. In comparing derivatives the differences in the fits become more obvious, but for the purpose of seeing a trend through time they are almost equal. It is clear from these graphs temperature is increasing over time. Some warmer weather occurs in the early 1950s, affecting the functions dramatically. Many climate sources explain the high temperatures of the 1950's as just variation of each year with no particular reason as to the specific cause[1]. Other scientists are more skeptical and say this could just be due to measurement error since recording temperatures was not regulated until later [4]. From other graphical representations, which implement different methods to visualize temperatures of the US throughout the 20th century, this peak is minor. The effect is amplified with functional analysis because spline functions are less accurate around the endpoints. The functions detect these higher values and don't ignore them as outliers. Though, even with this overall warmer temperature trend in the beginning of the data set, it is still discernable that the temperatures in the 2010s are higher than all other times in this data. The IPCC has reliable information on temperature dating back as far as the 19th century and no temperatures during any year compare with the high temperatures of the 2010s[8]. This affirms the belief that temperatures have been consistently becoming warmer and there is no indication this course will change.



(a) Order 6 splines of the data of all cities with seasonal variation removed.

(b) Closer look at one of the lines on the graph on the right: only includes Burlington VT.

Figure 6: Order 6 splines of the data over the 64 years with seasonal variation taken out and using only 10 basis functions.

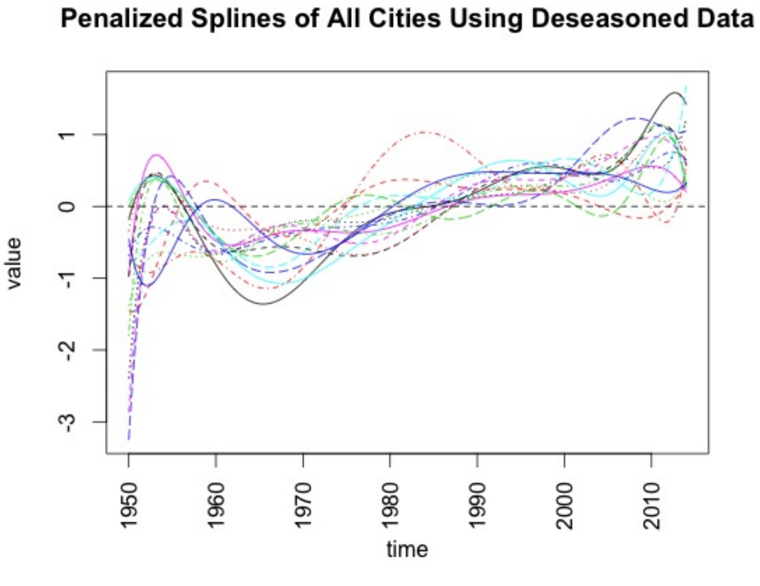


Figure 7: Splines of all the cities with seasonal variation removed penalized against the 4th derivative using 10^{10} as λ, the smoothing parameter, and 10 basis functions.

Although a glimpse of warming has been seen through the example of Burlington VT, it is also possible to look at the average of all of the cities together. The functional data objects were averaged together the same way any data set would be using the sum of the functions and dividing by 16 (number of functions contributing to the mean). 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 8. It is evident that even the upper limit of the 1950s is smaller than the lower limit of the 2010s, showing a definite upward escalation of temperature. This also gives us an indication of how the temperature is changing. The function is not a straight line so this pattern of upward momentum is not fully linear. Within most decades, the shape of the function changes and it is possible to see that the warming isn't linear. An example of this can be seen between years 1950 and 1960. There is a high peak but then it seems to jump back down to previous lower temperatures. This shows natural variation within the overall trend and can help explain why even in todays climate, winters can still see record low temperatures.

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. These calulations were based on the sampling distribution of the coefficients of the basis functions of the spline used on the data with seasonal variation removed. Figure 9 depicits the true confidence limits for one specific city, Burlington VT. The high degree of precision evident from the plot is partially the result of the removal of significant noise earlier in the analysis process. Figure 8 and 9 both present a significant observed difference in temperature from the start to finish. Upon further inspection Figure 9 shows Burlington was one of the cities whose temperature spiked in the 1950s and then dramatically drop to pre-50s levels of temperature and eventually began to consistently rise again. For Burlington, it also appears during the early part of the new millenium the temperature did not change much and may have even slightly decreased. But around 2005 the temperature quickly rose again and hit around one degree Celsius above the average of these years. This finding is consistent with many other studies stating 2013 was one of the warmest recorded years globally and we can see this on even a single city[11].

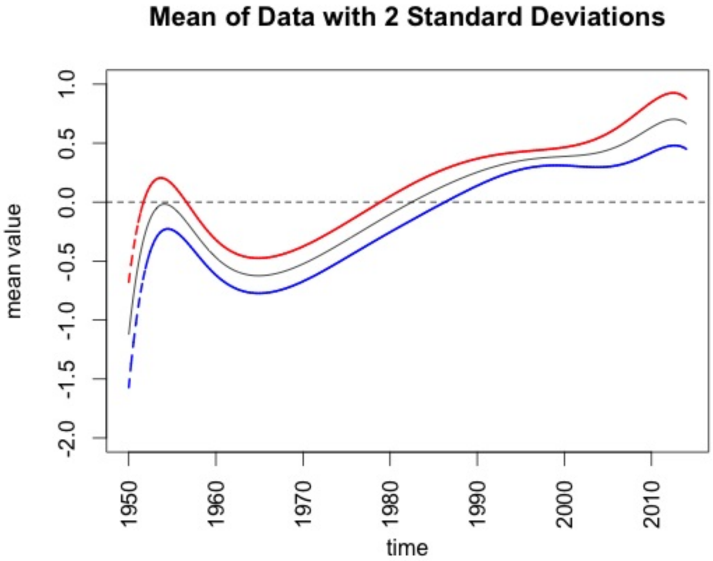


Figure 8: Mean of the penalized splines with seasonal variation removed and using 10^{10} as λ. The bands surrounding the mean are 2 standard deviation units from the mean computed by using the common sample standard deviation calculation.

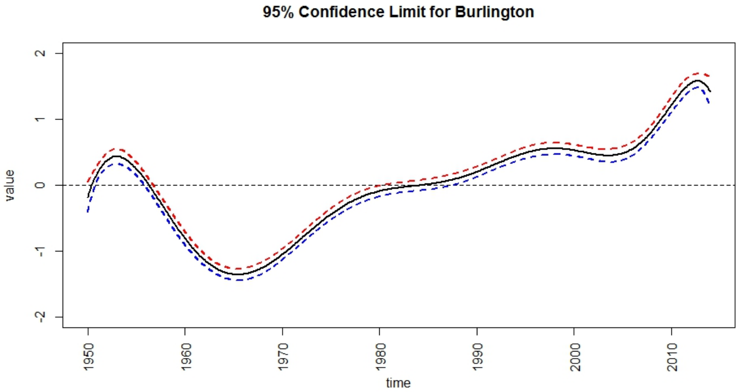
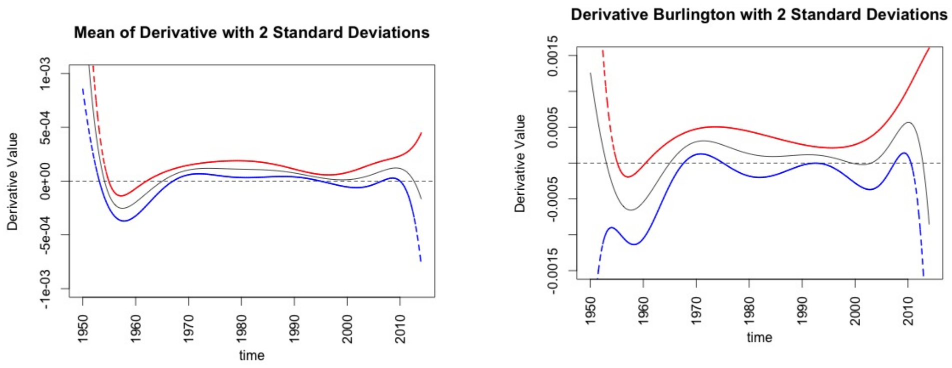


Figure 9: 95% Confindence Limit for the city of Burlington VT. The standard deviation used was calculated based on the coeffecients of the spline basis functions rather than considering the functions as observed data.

To further understand the temperature trend of the last 64 years, an investigation into derivative estimates was used. Figure 10 (a) depicts the mean of the derivative functions and Figure 10 (b) examines only the city of Burlington VT. The scale of the derivative is very small, on the order of 10^{-3} because it measures the change of temperature per day. Most studies report the average temperature over 1950 though 2010 is at most one degree Celsius [4] and in our case that change is distributed over more than twenty thousand observations. Similar to Figure 8, two times the standard deviation of the derivative mean was used as limits for the mean function. From the late 1960s until the early 1990s, even the lower limit was above the zero mark meaning temperatures were in fact increasing at a significant rate. For the derivatives, it is best to not make inferences near the endpoints because the splines are very extreme on the edges when referencing the derivative[5]. The high temperatures in the 1950s created the substantial negative derivative at the beginning of the graph because the variation during that time was so different compared to the years surrounding it. When a single city (e.g. Burlington VT) is examined in isolation, the standard deviation increases resulting in less precision for estimating effects over time.



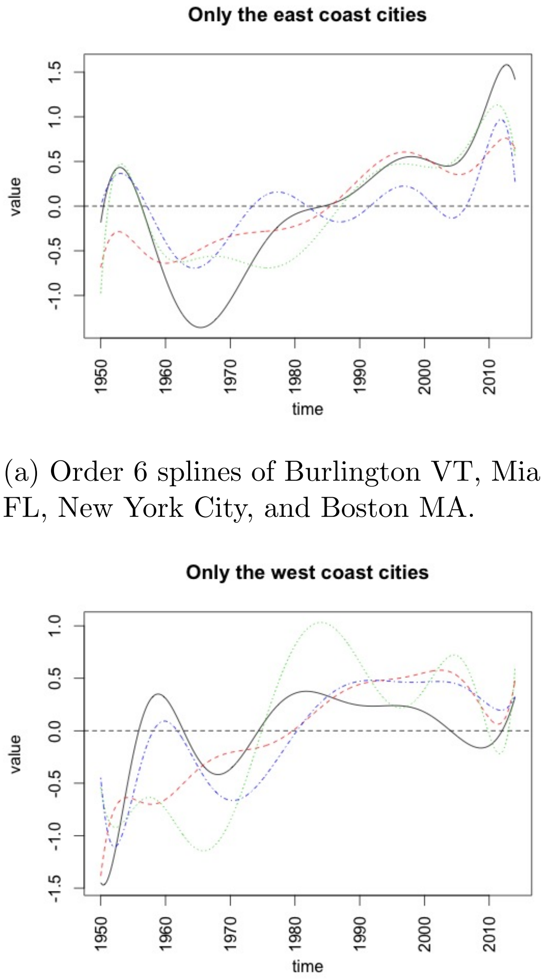
(a) Mean of derivative function with 2 standard deviations of the mean.

(b) Derivative function of only Burlington VT with 2 standard deviations.

Figure 10: First derivative of the penalized splines with seasonal variation removed and using 10^{10} as λ.

Finally, a permutation test was used to compare two groups of cities to determine if there was a difference between the trends of different regions' temperature. The groups of interest were cities of the East Coast versus cities of the West Coast. The Climate Central's data explores in detail the differences between each state and the increase of temperature per decade. Some states in the west and south only have gone up one tenth of a degree Celcius while others especially in the north east are estimated to have increased by a almost third of a degree every decade since the 1970s[9]. The hypothesis under consideration for this study is whether there is a significant difference the mean functional representation of temperature between the west and east coast.

Figure 11(a) represents the cities in this study, which are considered East Coast: New York City, Boston, Burlington, and Miami. Figure 11(b) displays Los Angeles, San Francisco, Portland, and Fairbanks, these were considered West Coast cities. From these figures there seems to be a big difference in the residuals during the 1950s and 1960s. Both coastal regions appear to be increasing together and to be above the zero average over time. A permutation test was used to calculate the overall significance level for the test and Figure 12 shows that the observed permutation statistic never crosses this threshold. A point wise level of significance was calculated and represents the significance level of each time value if a t-test was tested between the two groups at that specific location. We can expect there to be some time points were this is significant just because of random random error. Using pointwise significance there is a difference between the East and West Coast at a couple different times. There seems to be a difference between the 1950 and 1960 and also during the 2010s and it is also nearly significant in the permutation tests at those times. This is something to consider in future analyses when compiling the two coasts together to create a whole picture of the US. It could be that instead of treating them as a collective set they should be considered as distinct parts of a whole[9].


Only the east coast cities

(a) Order 6 splines of Burlington VT, Miami FL, New York City, and Boston MA.

Only the west coast cities

(b) Order 6 splines of Los Angeles CA, Portland OR, Fairbanks AK, and San Francisco.

Figure 11: East Coast and West Coast Cities with Season Variation Removed.

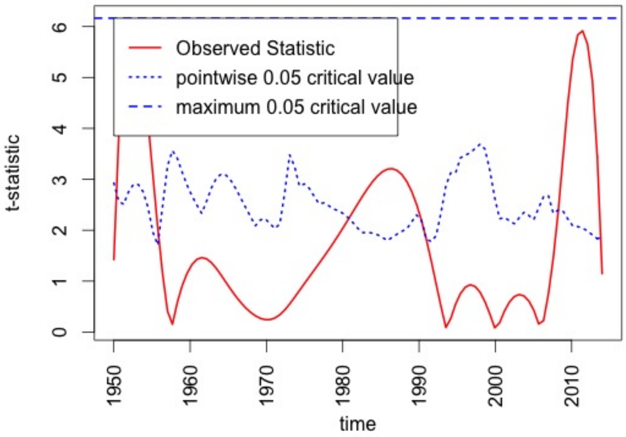


Figure 12: T permutation test comparing the spline functions without seasonal variation of the west and east coast cities in this study.

6 Discussion

This paper has described the purpose, concepts, and some of the methods of functional data analysis in the context of analyzing temperature trends. These methods were applied to 64 years of weather data in 16 cities in the United States to evaluate the trend of increasing temperatures the world is experiencing. Instead of working with discrete data, functional data analysis allowed for statistical tests to be performed on the data represented as functional objects.

The process of transforming the data into a functional form, while guided by objective criteria, is nonetheless somewhat of an art form. These criteria included generalized cross validation tests and minimization of sum of squared residuals.

The statistical tests on the weather functional data object provided insight into how the temperatures of US cities have increased significantly in the last six decades. The mean function of the data with a 95% confidence limit proved to be increasing significantly. There was no indication this upward trend of temperatures would stop, potentially leading to many serious unavoidable disasters unless the temperature increases stop.

This work has shown it is possible to detect changes in temperature and provided insight as to the rate of these changes. Different methods of analysis were explored along with different ways to determine knots, smoothing parameters, and other parameters specific to this data set. This paper did not explore some functional data methodologies, e.g. curve registration, a method still under development. A relatively small (in comparison to the new era of big data) data set was utilized that only included complete and evenly spaced data points. Further analysis could allow for incomplete data, which would allow a much longer span of time, with more cities, whose data may not be complete for every day, to be included.

References

[1] Climate Conservative Consumer.Extreme Climate Change Events: Early 1950's. From *http : //www.c3headlines.com.*

[2] Hall, P. and Keilegom, I. (2007). Two-sample tests in functional data analysis starting from discrete data, Statistica Sinica, 17, 1511-1531.

[3] Levitan, D., Nuzzo, R., Vines, B., and Ramsey, J. (2007). Introduction to functional data analysis. Canadian Psychology, 48(3), 135-155.

[4] Karl, T. R., Melillo, J. M., and Peterson, T.C. (eds.) (2009). Global Climate Change Impacts in the United States. Cambridge University Press. From *http : //downloads.globalchange.gov/usimpacts/pdfs/climate − impacts − report.pdf.*

[5] Ramsey, J., Hooker, G., and Graves, S. (2009). Functional data analysis with r and matlab. New York. NY: Springer.

[6] Ramsey, J., and Silverman, B. W. (2002). Applied functional data analysis: Methods and case studies. New York. NY: Springer.

[7] Ramsey, J., and Silverman, B. W. (2005). Functional data analysis. (2nd ed.). New York. NY: Springer.

[8] Stocker, T.F., D. Qin, G.-K.Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P.M. Midgley (eds.) (2013). Climate Change 2013: The physical Science Basis.Contribution of Working Group 1 to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.Cambridge University Press. From *http : //www.climatechange*2013*.org.*

[9] Tebaldi C., Adams-Smith, D., Heller, N. (2012). The Heat is On: U.S. Temperature Trends. Princteton NJ. From *www.climatecentral.org.*

[10] Thomas C.D., Cameron A., Green R.E., Bakkenes M., Beaumont L.J. et al. (2004). Extinction risk from climate change. Nature 427, 145-148.

[11] World Meteorological Organization (2014). WMO statement on the status of the global climate in 2013. From *http : //www.wmo.int/pages/index*$_e$*n.html.*