Functional Data Perspectives in GHCN Temperature Dataset

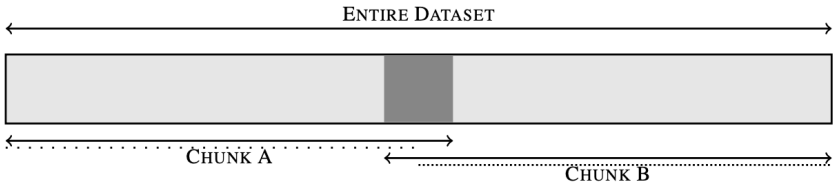
Zhiling Gu *

1 Overview

- **Reference Paper** Hadjipantelis, P., Müller, H.G. (2018). Functional Data Analysis for Big Data: A case study on California temperature trends. Handbook of Big Data Analytics Ed. Härdle, W., Lu, H.H.S., Shen, X., Springer Heidelberg, 457-483.
- **Dataset** GHCN (Global Historical Climatology Network).
  - **Dimension** 45990×4, each row is the daily observations for the maximum, the minmum and the spread of daily temperature from 1898 to 2019. Functional time series data.
  - **Issue** Original dataset exhibits high variance.
  - **Solution** Smooth the original dataset before performing analysis.
- **Objective** Find out if there is difference in temperature behaviours between Davis and Redwood City, and if there is a general temperature trend across time.
- **Conclusion** (1) Davis experiences in general higher variations in temperature which can be explained by larger mass of land. (2) There is an upward trend for daily minimum temperature, a downward trend for daily maximum temperature and the daily temperature spread.

2 Data Manipulation

- **Presmoothing** This dataset is in the form of functional time series data. We split the data annually and regard each of them as a realization of a function (with bounded domain) plus IID white noise. Under this setting, it is legitimate to perform smoothing to eliminate the white noise before analyzing the functional data. See Figure 1.
- **Computational remarks**
  - **Splitting** In R, the smoothing cannot be done once for all due to the limit of memory space. Splitting the data into q segments (chunks) and smoothing each of them are required.
  - **Edge Issue after Splitting** To alleviate the roughness on the edges between each segment, the author smoothed the data with extended segments at each ends of the segment, and evaluate only on the support of each segment. That means, the data of the first and last observations of each segment is used twice to provide information of the edge behaviour between two consective segments. The size of the overlapped segment is proposed to be $2bq$ where b is the bandwidth, where b is the bandwidth applied for smoothing. In this project, $b = 49$ days and Epanechnikov kernel is used.



*STAT547 Final Project, 2019 Fall, Iowa State University. Instructed by Prof. Xiontao Dai.

– **Smoothed Data** After pre-smoothing, we can work on the smoothed data. But be careful about the smoothed data, some of them can be infinite bacause missing observations occur at the beginning or the end of a year. In this project, I simply use the smoothed data on days with observations. The datasets and source code are available at GoogleDrive.

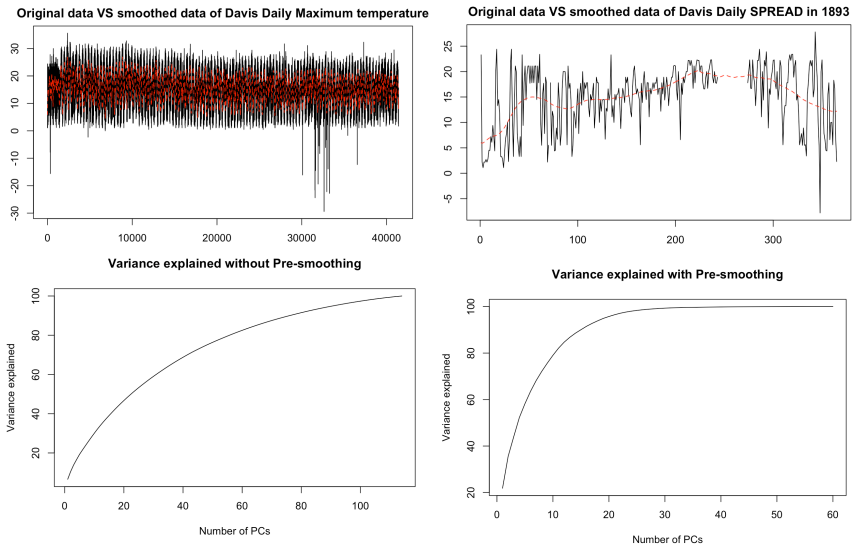


Figure 1: Pre-smoothing is essential for FPCA to provide meaningful explanation of the temperature behaviours.

3 Comparison between Redwood City and Davis

By looking at the mean and covariance functions estimates we can infer some differences in the temperature behaviour across the year. See Figure 3, Figure 4 and Figure 5 for details.

4 Trend Across Years

Apart from the previous discussion on the difference between two cities within a year, we are also interested in the overall trend of the temperature, i.e. if the temperature of a year, regarded as a function, in fact shifted structure-widely, i.e. each years temperatures no longer share the same mean function.

Shift in Mean

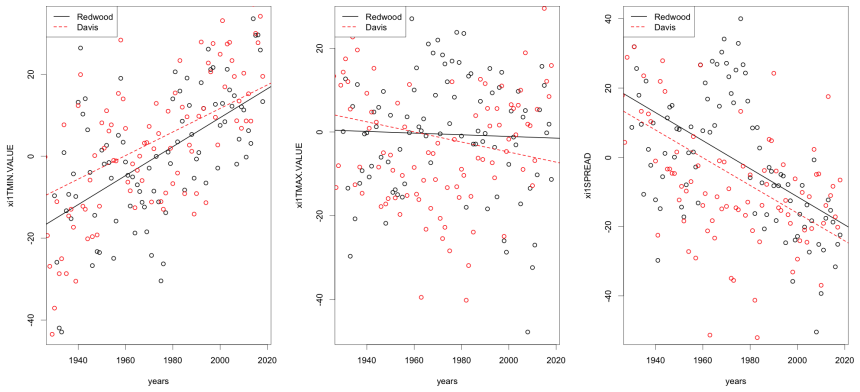


Figure 2: **Scatterplot of the first FPC score against year**

	Min	Max	Spread
		t-statistics	
Redwood City	6.445655	*-0.3390321*	-5.943052
Davis	5.777814	-2.2537011	-7.819670
		P-Value	
Redwood City	6.186967e-09	*0.73540248*	5.653790e-08
Davis	6.684611e-08	0.02612685	2.910134e-12

Table 1: **Test of the first FPC score $\hat{\xi}_1(t)$ across time (years).** The first FPC score represents the overall level of the temperature. Increase in the FPC score can be an indicator of increase in temperature. Except for the Redwood City daily maximum, we find all the other criterion shifted significantly across years. The daily minimum has an upward trend, the daily maximum has a downward trend, the daily variation (spread) has a downward trend.

Structure in Noise

On top the previous anlysis, the author also looked into whether the noise has structure against years. The functional variance process (FVP) is introduced to extract the information in the stochastic time-trends in the noise variance of functional data.

Consider $Y_1, \ldots, Y_n$ to be n continuous smooth random functions defined over a real interval $[0, T]$, Here we also assume that these functions are observed at a grid of dense time-points $t_j = \frac{j-1}{m-1}T, j = 1, \ldots, n$ with measurements

$$Y_{ij} = Y_i(t_j) + R_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m,$$

where R_{ij} are additive noises such that $E(R_{ij}R_{i'j}) = 0$ for $i \neq i'$ and $E(R) = 0$ and $E(R^2) < \infty$. Also we model the noise into two parts: $V(t_{ij})$ from the underlying smooth functional variance process, $W(t_{ij})$ is a noise component. We also assume these two component satisfies

$$R_{ij}^2 = \exp(V(t_{ij}) \exp(W(t_{ij})$$

Perform change of variables, we have

$$Z_{ij} := \log(R_{ij}^2) = V(t_{ij}) + W(t_{ij})$$

$$E(Z_{ij}) = E(V(t_{ij}) =: \mu_V(t_{ij})$$

$$Cov(Z_{ij}, Z_{i'j}) = Cov(V(t_{ij}), V(t_{i'j}))$$

Further assume mean and autocovariance of $V(t)$ as follows

$$E(V(t)) = \mu_V(t)$$

$$C_{VV}(s,t) = \sum_{k=1}^{\infty} \rho_k \Psi_k(s) \Psi_k(t)$$

where $\rho_1 \geq \rho_2 \geq \ldots, 0$ are non-negative ordered eigen values, Ψ_k being the corresponding orthonormal eigen functions of V, we are left with a pair of process

$$Y(t) = \mu_Y(t) + \sum_{k=1}^{\infty} \xi_k \phi_j(t)$$

$$V(t) = \mu_V(t) + \sum_{k=1}^{\infty} \zeta_k \Phi_k(t)$$

The author performed FPCA on the functional variance process and concluded there is a trend of reduction in variance process across years.

Appendix: Figures

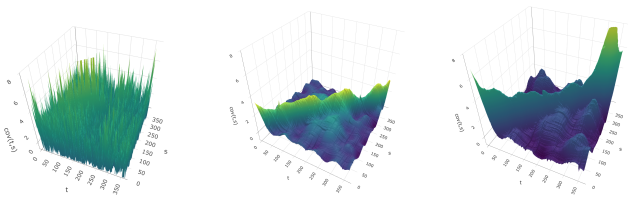


Figure 3: **Covariance function.** LHS: Estimated Covariance of Daily Temperature Spread (Davis). Middle: Estimated Covariance of Daily Temperature Spread (Redwood City). RHS: Estimated Covariance of Daily Temperature Spread (Davis).

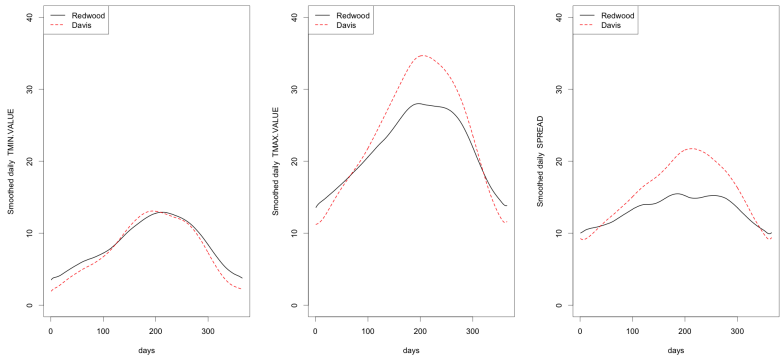


Figure 4: **Mean function.** Variation of temperatures is considerably higher in Davis. Since Redwood City is nearer to the sea where the temperature variation is lower.

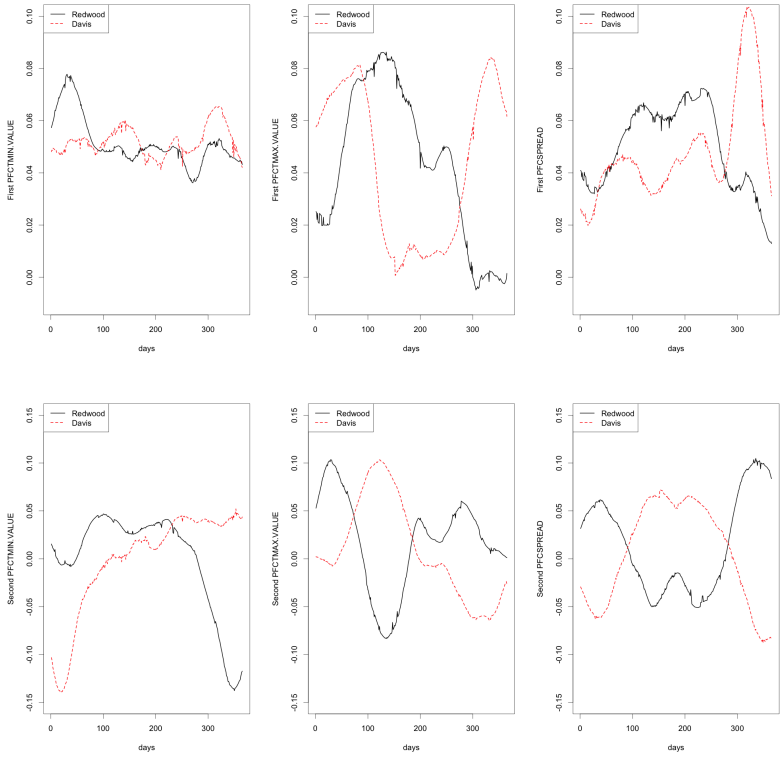


Figure 5: **First & Second eigenfunction**: The first eigen function shows the periods of highest variation of two cities do not coincide; The second eigenfunction settles in the neighborhood of zero, suggesting non-monotonicity for the temperarture and its variation throughout the year.