VILNIAUS UNIVERSITETAS

MATEMATIKOS IR INFORMATIKOS FAKULTETAS

MODELIAVIMO IR DUOMENŲ ANALIZĖS MAGISTRANTŪROS STUDIJŲ PROGRAMA

Magistro baigiamasis darbas

Funkcinių duomenų atvejo analizė

Functional Data Analysis - Case Study

Liyu Cao

Darbo vadovas: Alfredas Račkauskas

Vilnius, 2021

VILNIUS UNIVERSITY

FACULTY OF MATHEMATICS AND INFORMATICS

MODELLING AND DATA ANALYSIS MASTER'S STUDY PROGRAMME

Master's thesis

Functional Data Analysis - Case Study

Funkcinių duomenų atvejo analizė

Liyu Cao

Supervisor: Alfredas Račkauskas

Vilnius, 2021

Funkcinių duomenų atvejo analizė

Santrauka

Funkcinė duomenų analizė – tai informacijos apie tam tikrame intervale apibrėžtas funkcijas ar kreives analizė, sutrumpintai vadinama "FDA". Tai reiškia, kad stebimi duomenys turi būti interpretuojami kaip funkcijos ar kreivės, o ne duomenų taškų rinkinys.

Šio darbo tikslas yra ištirti kai kurias FDA metodikas, pradedant nuo pirmojo žingsnio – iš turimų duomenų atstatant imties kreivių formą. Tyrimas apima vieną iš populiariausių metodų, funkcinių pagrindinių komponenčių analizę (fPCA). FPCA yra populiarus ir naudingas įrankis dimensijų mažinimui. Be to, sukurtas funkcinės tiesinės regresijos modelis. Diskretieji kiekvieno dalyvaujančio kintamojo duomenys suglodinami naudojant Furjė ir B-splainų bazines funkcijas, panaudojus šiurkštumo baudą funkcijų glodumui nustatyti apibendrintos kros-validacijos (GCV) metodu. Galiausiai R^2 ir F-santykio funkcijos panaudotos nustatant funkcinio tiesinio modelio tinkamumą. FDA metodai yra taikomi klimato duomenų rinkiniams.

Raktiniai žodžiai : Funkcinė duomenų analizė; funkcinių pagrindinių komponentų analizė; funkcinis boxplot; fANOVA; funkcinė tiesinė regresija; klimato duomenys;

Functional Data Analysis - Case Study

Abstract

Functional Data Analysis refer to the analysis of information on functions or curves defined on some interval, which abbreviated as “FDA”. This involves thinking of the observed data as functions or curves rather than a set of data points.

The aim of this thesis is to investigate some FDA methodologies, starting with the very first step of constructing the functional form of the sample curves from their discrete data points. The study includes one of the most popular techniques of which is *Functional Principal Components Analysis*(fPCA). FPCA is a popular and useful tool for dimension reduction. In addition, a *Functional Linear Regression Model* is constructed. The discrete observed data points for each involved variables are smoothing by *Fourier Basis* and *B-spline Basis*, *roughness penalty* used to control the degree of the smoothness, smoothing parameter set by *generalized cross-validation* (GCV). Finally,the functions of R^2 and F-ratio were computed for the goodness-of-fit of functional linear model. Each of the aforementioned Functional Data Analysis methodologies applies to a real climate data set.

Key words : Functional data analysis; functional principal component analysis; functional boxplot; fANOVA; functional linear regression; climate data;

Contents

1 Introduction **3**
- 1.1 Literature Review . . . 3
- 1.2 Objective . . . 6
- 1.3 Layout of Thesis . . . 6

2 Methodology **7**
- 2.1 Smoothing Techniques . . . 7
- 2.2 Functional Descriptive Statistic . . . 11
- 2.3 Functional Principal Component Analysis . . . 12
  - 2.3.1 *Karhunen-Loeve* Expansion . . . 12
  - 2.3.2 Functional Principal Components Analysis . . . 14
  - 2.3.3 Varimax Rotation . . . 15
  - 2.3.4 Bagplot, Boxplot for Functional Data . . . 16

3 Functional Linear Model **18**
- 3.1 Functional Analysis of Variance (fANOVA) . . . 19
- 3.2 Fully Functional Linear Model . . . 21
- 3.3 Goodness of Fit [42] . . . 23

4 Applications: Case Study on Climate Data **24**
- 4.1 Variables . . . 24
- 4.2 Data Preparation . . . 25
- 4.3 Representing Functional Data . . . 28
  - 4.3.1 Temperature . . . 28
  - 4.3.2 Humidity . . . 31
  - 4.3.3 Pressure . . . 32
  - 4.3.4 Wind Direction . . . 33
  - 4.3.5 Wind Speed . . . 34
- 4.4 Functional Principal Component Analysis . . . 35
  - 4.4.1 Temperature . . . 35
  - 4.4.2 Humidity . . . 37
  - 4.4.3 Pressure . . . 40
  - 4.4.4 Wind Speed . . . 42
  - 4.4.5 Wind Direction . . . 45
- 4.5 Functional Linear Model . . . 47
  - 4.5.1 Functional analysis of variance . . . 47
  - 4.5.2 Analysis of Variance Model . . . 48
  - 4.5.3 Fully Functional Regression Model . . . 60

5 Conclusion **70**
- 5.1 Concluding Remarks . . . 70
- 5.2 Recommendations . . . 71

Appendices **77**

1 Introduction

Functional Data Analysis refers to a collection of methods for analysis of information on functions or curves defined on some interval, which abbreviated as "FDA" . Under an FDA framework, it treats each sample element as an individual realization of an underlying random process [40].

In traditional data analysis, data generally has such features that the data is either time series data or cross-sectional data. Many statistical data obtained in practice often take multiple cross-sections on the time series, and then select sample data constituted by sample observations on this interface at the same time. Such data called "panel data". Many researchers have studied the panel data, Panel data can alleviate the problem of insufficient sample size and has many advantages such as identifying the impact of factors that are difficult to measure. However, the econometric model of panel model uses a linear structure to describe the causal relationship between variables, and such model relies on many assumptions, so that the specific application of these methods and the type of data applicable have certain limitations. In practical, people often encounter a type of data with obvious functional characteristic (referred to as Functional Data). The term "Functional data analysis" was firstly proposed by Ramsay in a paper named "*When the Data are Functions*" [41] published in 1982. Ramsay pointed out that the modern sophisticated data collection hardware can acquire a series of functional data, for example, eletro-encephalography (EEG) records, functional magnetic resonance image (fMRI) data in medical diagnosis, speech signal, spatial data, and so on [41]. Functional data is a kind of data in the form of a function or curves and usually have a value at each time point. If the observed time points are very dense, these data will show a functional characteristic in the data space more obviously. For example, some climate data of weather stations in certain areas, various economic indexes, real-time transaction data of a stock in stock market and so on. In functional data, the observations are changed from a static to a dynamic conceptualized. Therefore, when presenting a continuous function using classical multivariate statistical methods, it will bring about many disadvantages. In classical statistical, we tend to sample the observations at a limited number of points. In one hand, information between sampling points is lost or the important information about the smooth functional behaviour of the generating process that underpin the data is ignored [16]. In another hand, it also suffers the number of covariance parameters is much larger than the samples [41].

The core feature of FDA is that the data is functional. The phenotypic form of functional data is a smooth curve or continuous function $x_i(t)$, $i = 1, 2, .., N$, where N is the number of observations, t typically represent a time variable, but can be any variable . "Function" refers to the internal structure of the data, not the external manifestation of the data. This thesis will present some methods for analysing such data with functional characteristic, a field known as Functional Data Analysis (FDA).

1.1 Literature Review

Ramsay and Dalzell (1991) [44] introduced some methods and tools suitable for functional data analysis involves infinite dimensions, and conducted functional principal components analysis and linear model on real data set of Canadian temperature and precip-

itation. J.O.Ramsay and B.W.Silverman (1997) summarized the theory and methods of functional data analysis and published the book "Functional Data Analysis"[43]. The book comprehensively expounds the basic characteristics of functional data and its idea of statistical analysis, by extending some traditional statistical methods to make them suitable for functional data, which greatly promotes the development of functional data analysis. Since the influential books by Ramsay and Silver([41], [44], [42], [43], [56], [45]), Frédéric Ferraty and Philippe Vieu ([13]), functional data analysis methods was popularized, and have been applied in many different fields, such as medical([2]), biology([23]), genomics([26]) and so on.

Ana M.Guilera et al.(2008) [2] applied the functional data analysis method to the medical field, using systemic lupus erythematosus patients as samples, and constructed a logistic regression model based on functional data, aiming to give the probability of a bivariate predictive dependent variable in terms of discrete time observations of a functional predictor variable. And a functional PCA based approach is proposed in this paper. Finally, the results are compared by different logit approaches, using a sample of Lupus patients. In the field of public health, Farah Yasmeen et al. [67] also applied functional time series model to age-specific breast cancer mortality and predict future breast cancer mortality rates by age separately for white and black women. Sarah J.Ratcliffe et al. (2002) [48] applied modified functional regression to periodically stimulated foetal heart rates represents an improvement than the best standard linear regression model.

In the filed of biology, functional data analysis has also been well applied. Takayoshi Ikeda et al. [23] applied the statistical technique of FDA to the time series analysis of plankton monitoring data. The main point of this analysis is to reveal patterns in the seasonal cycle to access interannual variability of variables. By curve registration and higher order derivatives using fit FDA curves, differences in the seasonal progress were seen. It's anticipated that the technique of FDA is a useful approach that can be applied to a wide variety of marine ecological data.

In demographics, functional data analysis methods have also been widely used. Rob J.Hyndman et al. [21] used a functional data for forecasting age-specific mortality and fertility rates observed over time. The results show that it achieves better forecasting than the other approaches to forecast mortality.

In signal and waves processing , functional data analysis methods also show strong application prospects. J.Lucero [29] studied a FDA algorithm for the time normalization of voice signals, which is more flexible than the previous dynamic programming approach. For different application, the FDA algorithm allows the use of different optimization criteria, like the weighted combination of derivatives of the wavelets or the time-dependent weight function to emphasize segment of the time interval. The resultant functions are smooth and differentiable, which can be used for further analysis. This algorithm might have a wide application in many different fields.

The use of functional data analysis approach in the environmental field draws more and more attention to the public (Siegfried Hörmann et al. [17]; Curceac Stelian et al. [8]; P.Z., Hadjipantelis et al.[40]). Gao H.Oliver et al.[35] used functional data analysis methods to model the dynamics of diurnal ozone and nitrogen oxides cycles taking into account the continuous nature of the photochemical system. Final representative summer diurnal ozone profiles are constructed using functional clustering. Norshadida Shaadan et al. [54] highlighted the advantages of using functional based methods for accessing

the comparing the behaviour of PM_{10} pollutant during extreme haze years in the state of Malaysia. This study also can be extended to the usage of functional depth method for outlier or abnormal behaviour detector. Meredith C. King et al. [24] presented a functional data analysis approach to analysing $PM_{2.5}$ variability and change over space and time, which allows for a better understanding of the temporal trends in nitrate and sulfate levels. This approach allows for complete profile prediction for sites or times without data and confirm existing findings and yield new insights about $PM_{2.5}$ variation.

At present, FDA covers a wide variety of statistical methodologies including functional principal component analysis, functional canonical components analysis, and functional linear model, functional clustering analysis and so on.

Functional principal component analysis treats variables as a form of a function, so is the sample covariance matrix. Juhyun Park et al. [36] did an in-depth study on principal component analysis and pointed out the analyzing functional data often leads to find common factors. FPCA is a useful tool can summarize and characteristics the random variables in a functional space. An alternative fPCA that produces directed components is proposed, which can obtain more information and easier to interpret. This approach is demonstrated with simulated examples and real data. Properties of some special cases are also established.

Joon Jin Song et al. [57] further developed the application of principal component analysis in medical and biological. They proposed a general methodological framework of a unified approach with conjunction with functional principal components analysis, and then clustered the gene expression of the time course, constructed the optimal number of bases functions in the smoothing step and fPCA using cross-validation technology. Finally compared the performance with some other popular classifications methods. Simulated data analysis and real data analysis are conducted.

The analysis of variance of functional data is for statistical analysis of the differences of some aspect of the objects. It is often to group the data by different geographical area or several categories. Each group is composed of many individuals. In functional data analysis, it is not only need to determine whether the specific effects of each group are zero, but also if these effects are significant at a special time t. Analogously to the classical statistical of variance analysis, most of the statistical machinery available for analysis of variance is readily applicable to this functional problem. for example, error sum-of-squares function, the squared multiple correlation function and the F-ratio function. But because these functions are related to value of t, the analysis method is different from the standard multivariate statistical method. By calculating the values of the two functions R^2 and F-ratio, then plot the functions, you can see the fit of the model from R^2. From the size of the F-ratio function and the given significance level, each individual can be determined if there is a difference between individuals.

Functional regression is an area of research and the approach depends on if the response or covariates are functional or scalar. Ramsay and Dalzell(1991) [44] consider a functional regression model where both dependent and independent variables are functions, and Ramsay and Silverman (2005) [42] considered its modelling strategy thereafter. They estimated the model by least square methods and assessing the fit of model considering the square correlation function, R^2. Hidetoshi Matsui et al. [32] proposed a functional regression model where multiple functional predictors and a functional response. They used the Gaussian Basis functions along with regularization techniques for

transfer discretely observed data to a smooth function, of which provides a useful instrument for that. They estimated the functional regression model by *Least Square*, *Maximum Likelihood* and *Penalized Maximum likelihood*. Four modified criteria are implemented for selection of regularization parameters, they are *Generalized Information Criterion*, *modified Akaike Information Criterion* and *Generalized Bayesian Information*.

1.2 Objective

The aim of this master thesis is entirely related to the modelling of Functional Linear Regression by means of Functional Data Analysis methodology. We explored some methods of FDA, such as *Functional Linear Models* and *Functional Principal Component Analysis*, also conducting exploratory data analysis by means of FDA methodology. In this thesis, we will use a real climate data set as the case for implementation.

1.3 Layout of Thesis

The layout of this thesis is to provide sufficient information regrading explore Functional Data Analysis by means of FDA methodology. The next chapters will cover the followings:

- **Chapter 2** introduced some techniques of functional data analysis.

  The first part of this Chapter is representing the functional data. Two basis functions are employed to construct the functional data, Fourier bases and B-spline bases. The roughness penalty is also introduced used to measure the roughness of curves and Generalized cross-validation used to deal with model selection. All these techniques will be used throughout this thesis.

  The second part of this Chapter will provide some functional descriptive statistics.

  The third part of this Chapter will focus on Mathematics of Functional Principle Component Analysis. This Chapter helps to clarify a stochastic process can be written as a linear combination of basis functions, which is call the *Karhunen - Loeve* expansion. Also, provide two graphical method to identify some information based on two ordering methods, Tukey's halfspace and kernel density estimate.

- **Chapter 3** introduced the theory of *Functional Linear Regression* Model. The first part is the functional analysis of variance that can be consider one special case of functional linear model where predictors are scalar and response are functions. The rest part of this Chapter is the functional linear model where the predictors and response are functions.

- **Chapter 4** Applying all the methodologies mentioned in previous sections to a real climate dataset. Finally to compare different climate indicator profile can predict temperature profile better based on fully functional linear model where univariate functional response and univariate predictor.

- **Chapter 5** Give the conclusion and recommendations that accumulated throughout the this thesis.

2 Methodology

In Statistic, it concerned with obtaining information from observation X_i,...,X_N. The observation X_n can be scalars, vectors or other objects. While for functional data analysis, it concerned with observations which are considered as functions or curves defined over set T.

In this section, we will provide some mathematical foundations of the FDA-methods used in this thesis. First of all, we will show how a function can be structured and smoothed from discrete observations(Section 2.1), then provide some descriptive statistic of functional data(Section 2.2). Section 2.3 focus on *Functional Principal Components Analysis*.

2.1 Smoothing Techniques

The common design assumption of FDA that we assume there exists a smooth function x giving rise to the observed data. But in practical, the samples are observed at finite data points. So functional data analysis begins with converting discrete data points to smooth varying functions by smoothing techniques. This emphasizes patterns in the data by minimizing short-term deviations because of observational errors, such as measurement errors and inherent system noise.

In a situation where the initial observations are typically can be represented as

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

where t represent the time steps and y_{ij} is the j-th observations of the ith sample function, and x_{ij} is a smooth function. The estimation x_i takes into account a finite-dimension functional space $F = span\{\phi_1, ..., \phi_k\}$,where ϕ_k is a set of functional building blocks and K is the number of basis functions, so that the smooth function x_i is defined as

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

.

The parameters c_{ik} are the coefficients related to k-th basis functions for the i-th function.

There are many ways to convert these observed N discrete data points to a function $X_i(t)$ with values $x_i(t_{ij})$ computable for any desired argument value t. The common way to achieve that is assuming the sample curves can presented by a set of basis of functions and to fit the basis coefficients using smoothing or interpolation [1].

If these observations are assumed to be errorless, then the process is called interpolation. But if they have some observational errors, then the process of forming functions from discrete data may call for smoothing [42]. The basic method to solve this problem is to select a set of basis functions $\phi_k(t), k = 1, 2, ...N$ that are mathematically independent of each other, then use the linear combination of basis functions to give the estimated value $\hat{x}(t)$ of the function xt, that is $\hat{x}(t) = \sum_{k=1}^{m} c_k\phi_k(t)$. The choice of basis function is very important for the estimation of the derivative, because $\frac{d^m\hat{x}(t)}{dt} = \sum_{k=1}^{m} c_k\frac{d^m\phi_k(t)}{dt}$. Hence,

not only should the estimation of the function being considered, but also estimation of the first or higher derivatives.

There exists many smoothing methods in the studies, including use of Fourier smoothing, regression splines, kernel smoothing, spline series, wavelet bases, roughness penalties and so on. Two of the most important parametric smoothing methods are employed in this thesis, Fourier basis functions and B-spline basis functions. The performance of these two smoothing approaches was studied via real climate data set.

Fourier basis A parametric method is known as a basis representation since we use a known basis to smooth the data. One best known basis system is Fourier basis. The Fourier basis series is

$$\phi_k(t) = \begin{cases} 1 & \text{k =1} \\ sin(\frac{k}{2}\omega t) & \text{k=2r, } r \in \mathbb{N}^+ \\ cos(\frac{k-1}{2}\omega t) & \text{k=2r+1, } r \in \mathbb{N}^+ \end{cases}$$

and $r = 1, ..., \frac{K-1}{2}$, where K is the number of basis functions. The frequency parameter ω is related to the period T by the relation $\omega = \frac{2\pi}{T}$. Fourier basis functions are periodic that arranged in successive sine/cosine pairs except for the first term. And noticed K must be an odd number ([35] pg 45-46). Fourier basis is useful when analysing periodical functions, for example, examine annual trends with seasonal variation [42].

The function vector $\phi(t) = (\phi_1(t), \phi_2(t), ..., \phi_k(t))'$ is said to be orthogonal if the value of t_j are equally spaced within interval T and the period is equal to the length of interval T. Because of that, the cross-product $\Phi'\Phi$ is diagonal and can be made equal to the identity by dividing the basis functions by suitable constants, $\sqrt{n}$ for $j = 0$ and $\sqrt{n/2}$ for all other j. The Fast Fourier Transform (FFT) algorithm can make it possible to compute all the coefficients efficiently, which is one of the reasons why this basis became well known [56] [11].

B-spline basis One common approach worth to mentioned is B-spline smoothing, which constructed from polynomial pieces joined at many knots.

Once the knots are given, B-spines can be estimated recursively for any degree of the polynomial by using a numerically stable algorithm [1] [10]. B-spline smoothing is the most popular smoothing technique used, presumably because of its simplicity and flexibility for tracking a wide range of non-parametric and semi-parametric modelling situations [60].

Consider the sample curves as the observations of a second order stochastic process $X = \{X(t) : t \in \mathbf{T}\}$ and the sample functions belong to the Hilbert space $L^2(T)$. For a set of K B-splines basis functions defined as [10]

$$\phi_{k,1}(t) = \begin{cases} 1 & \text{if } t_k \leq t < t_{k+1} \\ 0 & \text{Otherwise} \end{cases}$$

$$\phi_{k,m}(t) = \frac{t - \xi_k}{\xi_{k+m-1} - \xi_k}\phi_{k,m-1}(t) + \frac{\xi_{k+m} - t}{\xi_{k+m} - \xi_{k+1}}\phi_{k+i,m-1}(t)$$

m is the order of the B-spline, $m-1$ is the degree[10]. Particularly, if the denominators are 0, i.e. $\xi_{k+m-1} = \xi_k$ or $\xi_{k+m} = \xi_{k+1}$, the fraction should be evaluated as 0. [10], [9], [63] separately provides the study of spline functions from basic to higher mathematical

level. A very useful rule that the order of the spline basis to be at least two higher than the highest order derivative to be used [42]. So in this thesis, a cubic B-spline will be used in section of case study to fit the sample curves.

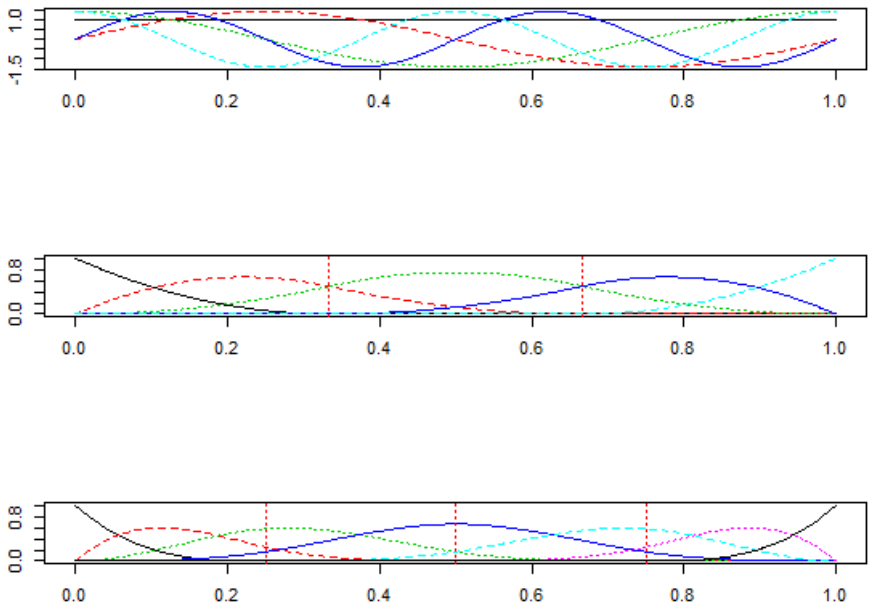


Figure 1. Simple example of the (a) Fourier series, and (b) (c) B-spline bases functions defined over the interval [0,1]. The order of B-spline basis are 3 and 4 with inter knot vector $\xi_{m=3} = [0, \frac{1}{3}, \frac{2}{3}, 1]$ and $\xi_{m=4} = [0, \frac{1}{4}, \frac{2}{4}, \frac{3}{4}, 1]$ respectively.

Next step is how to use these basis to fit a function to the data. Defined a matrix $\Phi(t)$ with dimensional $n \times K$ containing the elements $\phi_k(t_j)$:

$$\Phi(t) = \begin{bmatrix} \phi_1(t_0) & \phi_2(t_0) & \cdots & \phi_K(t_0) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(t_T) & \phi_2(t_T) & \cdots & \phi_K(t_T) \end{bmatrix}$$

in which column k represents the values of corresponding k-th basis functions evaluating at the data points of which has data for a subject. So the basis expansion function for $x(t)$ has the form:

$$\boldsymbol{y} = \Phi \boldsymbol{c} = \sum_{k=1}^{K} c_k \phi_k(t)$$

The solution of determining the coefficients vector $\boldsymbol{c}$ can obtained by minimizing the least square criterion:

$$SMSSE(\boldsymbol{y}|\boldsymbol{c}) = (\boldsymbol{y} - \Phi\boldsymbol{c})'(\boldsymbol{y} - \Phi\boldsymbol{c})$$

as described in [56](pg 59-60). By least square principle, taking the first derivative with respect to $\boldsymbol{c}$, we obtained the estimation of $\hat{\boldsymbol{c}}$ as:

$$\hat{\boldsymbol{c}} = (\boldsymbol{\Phi}'\boldsymbol{\Phi}^{-1})\boldsymbol{\Phi}'\boldsymbol{y}$$

Then get the fitted curves is:

$$\hat{\boldsymbol{y}} = \Phi\hat{\boldsymbol{c}} = \boldsymbol{\Phi}(\boldsymbol{\Phi}'\boldsymbol{\Phi}^{-1})\boldsymbol{\Phi}'\boldsymbol{y}$$

The Roughness Penalty For splines, the number of basis function K is defined by

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

Determining the number of basis functions K involves the selection of the number of knots. If too many knots are selected, it might cause overfitting of the data. On the contrary, if too few knots, you might suffer underfitting issues. One way to solve this problem is to select many knots and add a roughness penalty, which can trades off curve roughness against lack of data fit. This method restricts the flexibility of fitted curves and provides a good fit to the data with respect to the residual sum of squares and controls the degree of smoothness at the same time.

As stated before, the coefficient vector $\boldsymbol{c}$ can be estimated by minimizing the $(\Phi'\Phi)^{-1}\Phi'\boldsymbol{y}$. Simple least square approximation is suitable when we assumed the residuals are identically independent distribution with zero mean and constant variance. However, fitting basis expansions by least squares implies clumsy discontinuous control over the degree of smoothing [56]. Adding a *roughness penalty* term can solve this problem that could control the degree of the smoothness. The general version of the roughness penalized fitting criterion is

$$\min_{\mathbb{X}\in F} \sum_{j=1}^{n_i} \{y_{ij} - x_i(t_{ij})\}^2 + \lambda \int [Lx(t)]^2 dt$$

where $y_{ij} = x_i(t_{ij}) + \varepsilon_{ij}$ is the true value of observing x at time point t_{ij}, and true error $\varepsilon_{i,j}$ are statistically independent and have normal distribution with mean 0 and constant variance. If there are many knots, only minimizing the sum of squares errors is a least square issue, it will result in overfitting problem as it follows every details in the data. Hence, adding a roughness penalty term to the equation can solve this problem. The parameter λ emphasis on the second term penalizing the goodness of fit quantified in the sum of squared residuals in the first term.

There is a variety of ways to measure how "rough" or "wiggly" the curve is. One measure of a function's roughness is the total curvature, given by the integrated squared second derivative $\int [D^2x(t)]^2 dt$. In addition, Ramsay and Silverman(2005) introduced the differential operator $L = \omega^2 D + D^3$ as the harmonic acceleration operator, and the integral of the square of the harmonic acceleration operator may be an appropriate measure of roughness for periodic data like the temperature curves [42] [42]. The roughness penalty

method can produce a better estimation of derivatives than least square counterpart([56]). Hence, in the remainder of this thesis, the roughness penalty method is used.

Generalized Cross-Validation Referring to how to locate the penalty size, the generalized cross-validation measure (GCV) is popular developed by Craven and Wahba (1797). It is designed to find an optimal value of the smoothing parameter λ [42].

The GCV method comprises selecting λ so that the following criterion minimized:

$$GCV(\lambda) = (\frac{n}{n-df(\lambda)})(\frac{SSE}{n-df(\lambda)})$$

where $df(\lambda)$ measures the effective number of parameters estimating $x_i(t)$. In addition, the GCV method wins when comparing to k-fold cross-validation in the respect of computational, because GCV does not need to re-smooth the function k times as illustrated in Ramsay and Silver [42](2005, pg 97-99) .

Choosing a proper basis and its dimension for approximating functional form of sample curves is very essential. Ramsay and Silverman (2005 [56]) pointed up that the selection of smoothing technique is determined by the primary behaviour of the data being analysed [60]. In principle, the smooth curves should have featured the data being analysed. For instance, Fourier basis is generally selected when the data are periodic. There are many instances of the application using Fourier basis, environmental matters [35] [4] [31], cash flows process in Finance [27], and foetal heart rate tracing in the medical field [48]. Spline bases are regularly chosen to represent non-periodic data, like traffic volume forecasting ([62]) and EEG signal process ([70]) and wavelet bases are selected to represent data displaying discontinuities or rapid changes in behaviour [33].

2.2 Functional Descriptive Statistic

Once the functions or curves x_i fit to the data properly, some descriptive statistics can be analysed. In the classical statistic, the commonly used statistics for univariate data include mean, variance, covariance and correlation coefficient, and so on. Analogously, in functional data, when the observation is a smooth function $X(t)$ that is in L_2 and defined on the interval T with the following sample mean and variance functions:

$$\bar{x}(t) = (N-1)^{-1}\sum_{i=1}^{N} x_i(t)$$

and similarly:

$$\mathbf{var}_X(t) = (N-1)^{-1}\sum_{i=1}^{N}[x_i(t) - \bar{x}_i(t)]^2$$

where N is the number of sample functions or curves.

The covariance function encapsulate the dependence of records across different argument values, for all possible $t, s \in \mathbf{T}$:

$$\mathbf{cov}_X(s,t) = (N-1)^{-1}\sum_{i=1}^{N}[x_i(s) - \bar{x}(s)][x_i(t) - \bar{x}(t)].$$

and associated correlation function is:

$$corr_X(s,t) = \frac{\mathbf{cov}_X(s,t)}{\sqrt{\mathbf{var}_X(s)\mathbf{var}_X(t)}}$$

In our case, we have more than one observed variables, take pairs of functions (x_i, y_i) as an example, the cross-covariance function can be defined:

$$\mathbf{cov}_{X,Y}(t,s) = (N-1)^{-1}\sum_{i=1}^{N}[x_i(t)-\bar{x}(t)][y_i(s)-\bar{y}(s)].$$

Beyond above application, the functional standard deviation or the square root of the variance function **var** is concentrating on the underlying variability between observations. For example, the variations that are considered as measurement errors not attributable to the variability between observations [42] [**?**].

One more benefits using functional data analysis is derivatives. The derivatives of functional observations played an important role in functional data analysis. The first derivative of a function with respect to time gives a description of rate of change of variable, which corresponding to the velocity over time. The second derivative of a function with respect to time represent how the rate of change is itself changing, which corresponding to the acceleration over time. Plotting the first and second derivative as functions of argument, or plotting the second derivative as functions of the first derivative, may reveal important aspects of the processes generating the data. Ramsay and Silver (2005) introduced the idea of incorporating derivatives into a linear model for functional data and provide some applications (see [56], Chapter 17-19).

2.3 Functional Principal Component Analysis

2.3.1 *Karhunen-Loeve* Expansion

A basis function is a specific basis in function space and every continuous function can be represented as a linear combination of basis functions. In this section, we are going to use *Karhuen-Loeve* (K-L) expansion which is a presentation of a stochastic process in terms of an infinite linear combination of orthogonal functions. Given such kind of form of series to represent stochastic process was first considered by Kosambi (1943)[25]. In the situation of a centred square-integrable stochastic process $\{X(t)\}_{t\in[a,b]}$ over a probability space, if it satisfy a technical continuity condition, X_t can be represented as:

$$X_i(t) = \sum_{k=1}^{\infty}\xi_{ik}\phi_k(t)$$

where ϕ_k be the set of orthogonal basis functions in $L^2(a,b)$, ξ_{ik} are uncorrelated random coefficients. So, expansion if the process only taking K terms can be written as:

$$X_i(t) = \sum_{k=1}^{K}\xi_{ik}\phi_k(t)$$

It is use only a finite number of terms, K, leading to errors in the orthogonal basis decomposition, and K-L expansion is a representation of orthogonal basis decomposition that minimizes the mean square errors.

Consider a square-integrable stochastic process $\{X_i(t), t \in [a,b]\}$ with zero mean and continuous covariance function $K_X(s,t), t,s \in [a,b]$. The covariance function K_X satisfies the definition of a *Mercer kernel*, there exists a set $\{\lambda_k, \phi_k\}$ subject to

$$\int_a^b K_X(s,t)\phi_k(s)ds = \lambda_k\phi_k(t), (\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n \geq 0, n \to \infty)$$

.

where $\{\lambda_k, k \in \mathbf{N}\}$ are eigenvalues and $\{\phi_k, k \in \mathbf{N}\}$ is corresponding eigenfunctions. Further more, $\int_a^b \phi_k(t)^2 dt = 1$, $\int_a^b \phi_k(t)\phi_p(t)dt = 0 (k < p)$. The vital of K-L theorem is that it yields the best basis in the sense that it minimizes the total mean squared error [70]. Since the orthogonal basis functions used in K-L presentation are determined by the covariance function of the process, $\phi_k(t)$ is the corresponding orthogonal basis functions of covariance functions $K_X(s,t)$ with respect to k:th eigenvalues λ_k. And it's coefficient is

$$\xi_k = \int_a^b X(t)\phi_k(t)dt$$

and satisfies

$$E(\xi_k) = E[\int_a^b X(t)\phi_k(t)dt] = \int_a^b E[X(t)]\phi_k(t)dt = 0$$

$$\begin{aligned} E(\xi_i\xi_j) &= E[\int_a^b \int_a^b X(t)X(s)\phi_i(s)\phi_j(t)dtds] \\ &= \int_a^b \int_a^b E[X(t)X(s)]\phi_j(t)\phi_i(s)dtds \\ &= \int_a^b \phi_j(t)[\int_a^b K_X(s,t)\phi_i(s)ds]dt \\ &= \lambda_i \int_a^b \phi_j(t)dt \\ &= \delta_{ij}\lambda_i \end{aligned}$$

Using the fact that $\{\phi_j(t)\}$ is orthogonal, $\delta_{ij} = 0$ if $i \neq j$,$\delta_{ij} = 1$,if $i = j$. Hence, $\mathrm{Var}(\xi_k) = \delta_{kk}\lambda k = \lambda k$. The general case of a stochastic process X_i that is non-centralized can be brought back to the case of centred stochastic process by removing its mean $X_i - E[X_i]$.

Since

$$Var[X(t)] = \sum_{k=0}^{\infty} \phi_k(t)^2 Var(\xi_k) = \sum_{k=1}^{\infty} \lambda_k \phi_k(t)^2 = \sum_{k=1}^{\infty} \lambda_k$$

The total variance of the d-truncated approximation is $\sum_{k=1}^{K} \lambda_k$ and the d-truncated expansion explains $\frac{\sum_{k=1}^{K} \lambda_k}{\sum_{k=1}^{\infty} \lambda_k}$ of the variance.

2.3.2 Functional Principal Components Analysis

The principal components analysis (PCA) is a main methodology to reduce the dimensionality of a data set, in the meantime preserving as much statistical information as possible. In multivariate data analysis, this can be achieved by transforming to some new axis along which the data has largest variance. Denoting the projection of $\mathbf{X}$ on a new axis, principal components ξ_k, by

$$\boldsymbol{f}_k = \mathbf{X}\boldsymbol{\xi}_k$$

where $\boldsymbol{f}_k$ is the vector containing the k-th PC's scores, and PCA is trying to find sets of normalized weights $\boldsymbol{\xi}_k$ that maximize the variation in $\boldsymbol{f}_k$. The first PC can be described as $\boldsymbol{\xi}_1$ for the linear combination values

$$\boldsymbol{\xi}_1 = argmax[\boldsymbol{f}'\boldsymbol{f}] = argmax[\boldsymbol{\xi}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\xi}]$$

that have the largest variaation under the constraint of $\|\boldsymbol{\xi}_1\|^2 = 1$, where $\|\cdot\|$ represent the $L^2 - norm$. By the same manner, the subsequent components can be done by maximizing the mean square, subject to the constraint $\|\boldsymbol{\xi}_m\|^2 = 1$ alogn with the additional condition of orthogonality $\boldsymbol{\xi}_k'\boldsymbol{\xi}_m = 0,\ (m < k)$ [42]. Defined matrix $\mathbf{V}$ is the sample variance-covariance matrix $\boldsymbol{V} = (N-1)^{-1}\boldsymbol{X}'\boldsymbol{X}$. So the solution of maximization problem is solved by finding a sequence set of eigenvalue-eigenvector pairs(λ_j, ξ_j) satisfying the eigenequation $\boldsymbol{V}\boldsymbol{\xi} = \boldsymbol{\lambda}\boldsymbol{\xi}$.

The transition from classical multivariate data to functional data is not difficult, involving essentially replacing a summation by an integral[41]. In functional version of PCA, suppose we have a data $y_i(t)$ which can be considered as functional data. It's covariance function $v(s,t)$ defined as $N^{-1}y_i(s)'y_i(t)$. Analogous of multivariate statistic, the principal components' score are

$$\boldsymbol{f}_k = \boldsymbol{y}_i(t)\xi_k(t)dt$$

and the first weight function $\xi_1(t)$ are now

$$\xi_1(t) = argmax[\int_T \boldsymbol{y}_i(t)\xi(t)dt]'[\int_T \boldsymbol{y}_i(t)\xi(t)dt]$$

with the constraint $\int \xi_1(t)^2 dt = 1$. The subsequent weight functions ξ_m is required to satisfy $\int \xi_m(t)^2 dt = 1$ and additional orthogonality constraints that $\int \xi_k\xi_m = 0, m < k$. Each of the principal component weight function $\xi(t)$ satisfies the equation $\int v(s,t)\xi(t)dt = \lambda\xi(s)$. ([56] pg 150). Additionally, suppose that each function $y_i(t)$ and eigenfunction ξ have basis expansion that

$$y_i(t) = \sum_{k=1}^{K} c_{ik}\phi_k(t)$$

and

$$\xi(t) = \sum_{k=1}^{K} b_k\phi_k(t).$$

Write these equations by matrix notation as $\boldsymbol{x} = \boldsymbol{C\phi}$ and $\xi(t) = \phi(t)'\boldsymbol{b}$. Hence, to get the principal components, we require to find eigenvalue and eigenfunction pairs$(\lambda, \boldsymbol{u})$ that satisfying the following matrix:

$$N^{-1}\boldsymbol{W}^{1/2}C'C\boldsymbol{W}^{1/2},$$

and $\boldsymbol{W}$ is a order K symmetric matrix having entries $\int \phi\phi'$, principal component function's coefficient vector $\boldsymbol{b} = \boldsymbol{W}^{-1/2}\boldsymbol{u}$, where $\boldsymbol{u}$ is the eigenfunctions. (Ramsay and Silverman, 2005, pg 160-163 [42])

As stated in [40], functional principal components analysis(fPCA) is an intrinsically linear and unsupervised method to reduce dimension, which facilitates the conversion to infinite-dimensional functional data onto a finite-dimensional vector of random scores. Linear means the resulting representation is linear in the random functions.

In one sample of identically independent distribution smooth random functions $X_i \in L_2[T], i = 1,2,...n$, we assume a well-defined mean function is $\mu = E(X)$ and covariance function is $K_X(t,s) = cov\{X(t),X(s)\} = E[\{X_i(t) - \mu(t)\}\{X_i(s) - \mu(s)\}]$. Noted that $K_X(s,t)$ is symmetric positive definite, by Mercer's theorem, K has the representation:

$$K_X(s,t) = \sum_{k=1}^{\infty} \lambda_k\phi_k(s)\phi_k(t)$$

where $\lambda_1 \geq \lambda_2 \geq ... \geq 0$ are eigenvalues in decreasing order and ϕ_k's are the corresponding orthonormal eigenfunctions. So the K-L expansion of the observation X is given as:

$$X_i(t) = \mu(t) + \sum_{k=1}^{K} \xi_{ik}\phi_k(t)$$

where the ξ_{ik} are the fPC's scores with variance equal to the corresponding eigenvalues λ_k and eigenfunctions ϕ_k are the orthonormal basis by the random process X_i.

2.3.3 Varimax Rotation

Varimax rotation method can show more interesting properties of principal component analysis by rotating the matrix $\boldsymbol{\xi}(t)$ by a $K \times K$ orthnormal matrix R, such that

$$\zeta = \mathbf{R}\xi$$

.

Matrix $\mathbf{R}$ is chosen by maximizing the variance of the squared elements in ζ. $\mathbf{R}$ is a rotation matrix, the overall sum of variance is still the same. Therefore, if the values of elements are tend to be relatively larger or small , giving the PCs with more accentuated features. Also, after rotation, rotated component score are no longer uncorrelated [56][42].

2.3.4 Bagplot, Boxplot for Functional Data

Graphical methods could help to discover some data features that might not have been clearly visible using mathematical models [55]. Ramsay and Silverman ([45]) introduced the phase-plane plot, a plot of acceleration against velocity, which can highlight important distributional characteristic. So, graphical methods are the way to identify some information, which can not be obviously seen from a plot of original data. Hyndman and Shang (2010) [22] applied two useful ordering methods to the first two functional principal component scores([22]). These two graphical methods can work for outliers detection capability.

One of them is plot the functional data ordered by Tukey's halfspace depth (Tukey, 1975 [59]). Tukey (1975) proposed functional data $X = x_i, i = 1, ..., n$ can be ordered by halfspace location depth. The halfspace depth of an arbitrary point θ relative to a bivariate data set $Z = z_1, ..., z_n$ is given by the smallest number of data points contained in a closed half-plane, of which the boundary line passes through θ (Tukey,1977), where $\theta \in \mathbf{R}^2$ is one data point in $\mathbf{Z}$, then [55]

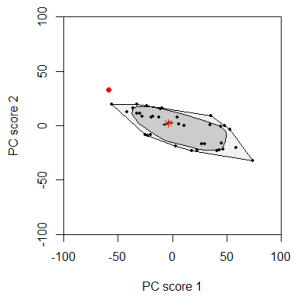
$$d(\theta, \mathbf{Z}) = \min_{H} \#\{i; z_i \in \mathbf{H}\}$$

.

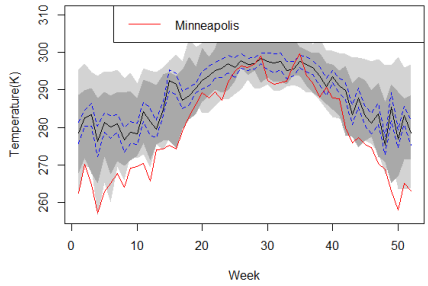
An element is close to the center of the sample will have a higher depth, while if it far away from the center will have a low depth. Given a fixed $\alpha \geq 0$, Rousseeuw and Ruts (1996)[50] defined the depth region D_α, the smallest number of data point in D_α is k,then

$$D_\alpha = \{x \in \mathbf{R}^2;\ d(x, X) \geq k\}$$

In year 1999, Rousseeuw, Ruts, and Tukey ([51]) proposed the bivariate bagplot to refer to the halfspace location. The robust bivariate principal component scores can be ordered by Tukey's halfspace location depth and plotted in a two-dimensional graph.



(i) The bivariate bagplot of smoothed temperature data

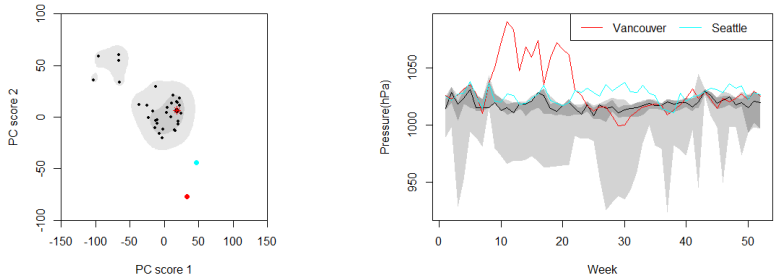


(ii) The functional bagplot of smoothed temperature data.

Figure 2. Bivariate bagplot and functional bagplot for the weekly temperature data of 36 cities. The dark and light gray region show the bag and fence regions, respectively. The deepest depth represented by the red asterisk mark. In the functional bagplot, the solid black line represents the median curve, surrounded by 95% pointwise confidence intervals. The curves outside the railing are outliers. The outlier city is Minneapolis.

Figure 2 is the example of bagplot. The bag colored in light gray defined as the smallest depth region in the sense that 50% of the total observations is falling into this region. The outer region obtained by inflating the bag by a factor ρ. In order to ensure that 99% of the observations fall into the fence that are projected bivariate scores follow standard normal distribution, ρ prefer being 2.58 [22]. The functional curves is another form of bivariate bagplot, exhibiting the median curve with the deepest depth, the inner and outer regions. Hence, 50% of functions are in the inner region. The bivariate outliers and functional outliers are lined by the same color.

Another one is ordering the curves by the value of kernel density estimate [53]. Let $o_i = \hat{f}(\mathbf{z}_i)$, $\hat{f}(\mathbf{z}_i)$ is a bivariate kernel density estimation which is calculated based on the bivariate principal component scores. Then order the observations by the value of o_i in a decreasing order. Hence, the first observations is the curve with the highest density and it may be treated as the "modal curve". Whereas the last curve has the lowest density value which may be treated as the most unusual curve. Hyndman (1996) [20]) introduced the bivariate boxplot to refer to the highest density, and mapping the features of the HDR boxplot into the functional space. Hyndman and Shang (2010) [22] compared these two new outlier detection methods with existing outlier detecting methods, and the results shows that the functional bagplot and boxplots are better to identify the outliers. One examples is shown in Figure 3.



(i) The bivariate HDR boxplot (ii) The functional HDR boxplot

Figure 3. The bivariate and functional HDR boxplot. The dark and light gray regions show the 50% HDR and the 95% outer HDR, respectively. These correspond directly to the equivalent regions with similar colours and shading in the right panel functional HDR boxplot. Points outside these outer regions are identified as the outliers.

3 Functional Linear Model

One application of FDA involves the construction of functional models that allow to explore the relation between functional dependent variable and functional explanatory variable [60]. Such models are called functional linear model. Functional linear model are the functional expansion of the linear models in multivariate statistics. Linear model can be functional in one or both of two ways:

- The dependent variable x is functional.
- One or more of the independent variable or covariates z is functional.

That means constructing the functional linear model depends on which variable being functional. Horváth et al. [18] proposed three prototypes of linear model with assumptions of the dependent variable and independent variables have mean zero, and errors ε_i are independent of the explanatory variables X_i.

The first scenario is:

$$Y_i(t) = \int \Psi(t,s)X_i(s)ds + \varepsilon_i(t)$$

in which both Y_i and X_i are curves that is called fully functional linear model.

Second scenario is:

$$Y_i = \int \Psi(s)X_i(s)ds + \varepsilon_i$$

when the regressor X_i are curves, but responseY_i are scalars, which is called scalar response model.

The other scenario is the functional response model:

$$Y_i(t) = \Psi(t)x_i + \varepsilon_i(t)$$

here the response Y_i are curves, while regressors x_i are scalars (Horváth, Lajos and Kokoszka, Piotr, 2012 [18]).

Every variable will have fitted through some bases as described in previous sections. The key problem is that the functions $\Psi(t,s)$ are infinite dimensional objects which must be estimated from a finite sample. $\Psi(t,s)$ reflects the effect of the explanatory function X_i at time s on the response function Y_i at time t.

3.1 Functional Analysis of Variance (fANOVA)

Functional analysis of variance extends from classical analysis of variance to functional data, abbreviated fANOVA. The aim at one-way fANOVA testing problem is to check if a factor or a category variable has a statistically significant effect. This factor is usually used to divide the individual functions into several groups [69]. Suppose we have k independent sample groups of functional data $X_{ij}(t), i = 1,2,...,k, j = 1,...,n_i$ defined over a given finite interval $T = [a,b]$. Let $\mathbf{Sp}(\mu,\gamma)$ denotes a stochastic process with mean function $E(X_i(t)) = \mu_i(t)$ for all $t \in \mathbf{T}$ and covariance function $\gamma(s,t)$, for all $s,t \in \mathbf{T}$. In FDA, t typically represents a time variable, $X_i(t)$ represents the smoothed curves. Functional One-way ANOVA problem is to test the main-effect functions are the same.

So the null hypothesis is :

$$H_0 : \mu_1(t) = ... = \mu_k(t),\ t \in \mathbf{T}.$$

the alternative hypothesis that its negation holds. If the one-way ANOVA model is not statistically significant, the null hypothesis is rejected. Then further investigation might be required. Post Hoc Test can be used to test if any two main-effect functions are the same[69]. This test can be written as:

$$H_O : \mu_i(t) = \mu_j(t), t \in \mathbf{T}$$

$$H_1 : \mu_i(t) \neq \mu_j(t), for\ some\ t \in \mathbf{T}.$$

Suppose $y_{ij}(t)$ is the j-th function under the influence of the i-th group ($i = 1,2,...k$; $j = 1,2,...,n_i$), then the model can be established as:

$$y_{ij}(t) = \mu(t) + \alpha_i(t) + \varepsilon_{ij}(t)$$

where the function $\mu(t)$ is the grand mean function, $\alpha_i(t)$ are specific effects on $y(t)$ in the factor of i. In order to identify such specific effects of different factors, we need to add the constraint $\sum_{i=1}^{k} \alpha_i(t) = 0$ for all t. The residual function ε_{ij} is the unexplained

variation of the j-th within group i. Defined regression function β_j by setting $\beta_1 = \mu$, $\beta_2 = \alpha_1$, ..., $\beta_{k-1} = \alpha_k$. So, we have the equivalent model formulation:

$$y_{ij}(t) = \sum_{g=1}^{k+1} z_{(ij)g}\beta_g(t) + \varepsilon_{ij}(t)$$

It also can be written in matrix notation:

$$\mathbf{Y}(t) = \mathbf{Z}\boldsymbol{\beta}(t) + \boldsymbol{\varepsilon}(t)$$

Where

$$\begin{aligned}\boldsymbol{\beta}(t) &= (\beta_1(t), \beta_2(t), ..., \beta_{k+1}(t))^T \\ &= (\mu(t), \alpha_1(t), \alpha_2(t), ..., \alpha_k(t))^T\end{aligned}$$

$$Z_{(kn_i)(k+1)} = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & & \\ 1 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}$$

Using the least square principle, our aim is to minimize the residual sum of squares. Hence, the least squares fitting criterion becomes:

$$LMSSE(\beta) = \sum_i \sum_j \int [y_{ij}(t) - \mu(t) - \alpha_i(t)]^2 dt$$

with the constraint $\sum_{i=1}^{k} \alpha_i(t) = 0$ for all t.

The sum of squares functions are a vital source of information if considering the importance of the different factor effects:

$$SSE(t) = \sum_{ij} [y_{ij}(t) - \hat{\mu}(t) - \hat{\alpha}_i(t)]^2$$

and the error sum of squares functions that taking only estimated grand mean function $\hat{\mu}$ into account:

$$SSY(t) = \sum_{ij} [y_{ij}(t) - \hat{\mu}(t)]^2$$

Then, F-test statistic and the squared multiple correction function RSQ are given by the following formula:

$$F - Ratio = \frac{\frac{(SSY - SSE)}{df(ref)}}{\frac{SSE}{df(error)}}$$

$$RSQ(t) = \frac{SSY(t) - SSE(t)}{SSY(t)}$$

.

where $df(error)$ is the degrees of freedom for error and $df(reg)$ is the number of mathematically independent functions β in the model ([69] [56]).

The roughness penalties approach mentioned in the previous section also can play an essential role in the *functional linear regression* model. As Ramsay et.al stated in [42], adding the roughness penalty can reduce the possibilities of missing important features and possessing extraneous features. Suppose parameter $\hat{\beta}$ can be expressed as $\hat{\beta} = \mathbf{B}\theta$, observed function y_i can be expressed as $\mathbf{y}(t) = \mathbf{C}\phi(t)$. Using a linear differential operator L to defined the roughness penalty for β, the penalized least squares criterion becomes:

$$PENSSE(y|\beta) = \int (\mathbf{C}\phi - \mathbf{ZB}\theta)'(\mathbf{C}\phi - \mathbf{ZB}\theta) + \lambda \int (\mathbf{LB}\theta)'(\mathbf{LB}\theta)$$

Smoothing parameter can be chosen by GCV method.

3.2 Fully Functional Linear Model

Fully Functional Linear Regression model constructed when the response and covariates are curves (functions). Let $Y_i(t), i = 1,..,N,\ t \in T$ are functional response and $X_{ik}(t), k = 1,..K$. K is the number of functional independent variables. In general, the model can be written as:

$$Y_i(t) = \beta_0(t) + \sum_{k=1}^{K} \int_{T_k} X_{ik}(s)\beta_k(s,t)ds + \varepsilon_i(t),\ s \in T_k,\ t \in T \tag{1}$$

The function $\beta_0(t)$ is the intercept function that captures the variation in the response that does not depend on any of the covariate functions. The bivariate regression coefficient function $\beta_k(s,t)$ reflects the effect of covariate functions $X_{ik}(s)$ on dependent function Y_i at each time t. $\varepsilon_i(t)$ are the error functions. By smoothing techniques, the functions $X_{ik}(t)$ and $Y_i(t)$ can be expand as a linear combination of some bases. Assume $\{\boldsymbol{\phi}_m,\ m \geq 1\}$ and $\{\boldsymbol{\psi}_g,\ g \geq 1\}$ are some bases such that

$$X_{ik}(s) = \sum_{m=1}^{M_k} \tilde{\eta}_{ikm}\phi_{km}(s) = \tilde{\boldsymbol{\eta}}'_{ik}\boldsymbol{\phi}_k(\boldsymbol{s}),\ s \in T_k$$

$$Y_i(t) = \sum_{g=1}^{G} \tilde{\theta}_{ig}\psi_g(t) = \tilde{\boldsymbol{\theta}}'_i\boldsymbol{\psi}(\boldsymbol{t}),\ t \in T$$

Consider the coefficient functions $\beta_k(s,t)$ has following form with double expansion:

$$\begin{aligned}\beta_k(s,t) &= \sum_{m,g} b_{kmg}\phi_{km}(s)\psi_g(t) \\ &= \boldsymbol{\phi}'_k(s)\mathbf{B}_k\boldsymbol{\psi}(t)\end{aligned}$$

where $\mathbf{B}$ is a $M_k \times G$ matrix of coefficients b_{kmg}. By centring the model (1) in the following way:

$$
\begin{aligned}
X_{ik}^*(s) &= X_{ik}(s) - \bar{X}_{ik}(s) \\
&= \tilde{\boldsymbol{\eta}}_{ik}'\boldsymbol{\phi}(s) - \bar{\boldsymbol{\eta}}_{ik}'\boldsymbol{\phi}(s) \\
&= \boldsymbol{\eta}_{ik}'\boldsymbol{\phi}(s)
\end{aligned}
$$

$$
\begin{aligned}
Y_i^*(t) &= Y_i(t) - \bar{Y}_i(t) \\
&= \tilde{\boldsymbol{\theta}}_i'\boldsymbol{\psi}(t) - \bar{\boldsymbol{\theta}}_i'\boldsymbol{\psi}(t) \\
&= \boldsymbol{\theta}_i'\boldsymbol{\psi}(t)
\end{aligned}
$$

Then, the model (1) becomes:

$$
Y_i^*(t) = \sum_{k=1}^{K} \int_{T_k} X_{ik}^*(s)\beta_k(s,t)ds + \varepsilon_i^*(t),\ s \in T_k,\ t \in T \tag{2}
$$

So, we have the following form of model:

$$
\begin{aligned}
\boldsymbol{\theta}_i'\psi(t) &= \sum_{k=1}^{K} \int_{T_k} \boldsymbol{\eta}_{ik}'\boldsymbol{\phi}(s)\boldsymbol{\phi}_k'(s)\mathbf{B}_k\boldsymbol{\psi}(t)ds + \varepsilon_i^*(t) \\
&= \sum_{k=1}^{K} \boldsymbol{\eta}_{ik}'\mathbf{J}_k\mathbf{B}_k\boldsymbol{\psi}(t) + \varepsilon_i^*(t) \\
&= \mathbf{z}_i'\mathbf{B}\boldsymbol{\psi}(t) + \varepsilon_i^*(t)
\end{aligned} \tag{3}
$$

where $\mathbf{J}_{\phi_k} = \int \phi(s)\phi(s)'ds$ is a $M_k \times M_k$ matrix, $\mathbf{z}_i = (\eta_{i1}'\mathbf{J}_{\phi_1}, ..., \eta_{iK}'\mathbf{J}_{\phi_K})'$ which is a vector with length $\sum_{k=1}^{K} M_k$, and $\mathbf{B} = (\mathbf{B}_1, ..., \mathbf{B}_K)'$ is a matrix with dimension $(\sum_{k=1}^{K} M_k \times G)$. Combining all the information above, we obtain:

$$
\mathbf{D}\boldsymbol{\psi}(t) = \mathbf{Z}\mathbf{B}\boldsymbol{\psi}(t) + \boldsymbol{\varepsilon}(t) \tag{4}
$$

where $\mathbf{D}$ is $N \times G$ matrix, $\mathbf{Z}$ is a $N \times (\sum_{k=1}^{K} M_k)$.

$\mathbf{B}$ in above model (4) can be estimated by minimizing the integrated sum of squares [56]:

$$
\begin{aligned}
&\sum_{i=1}^{N} \int [Y_i^*(t) - \sum_{k=1}^{K} \int_{T_k} X_{ik}^*(s)\beta_k(s,t)ds]^2 dt \\
&= \int_T \{(\mathbf{D}\boldsymbol{\psi}(t) - \mathbf{Z}\mathbf{B}\boldsymbol{\psi}(t))(\mathbf{D}\boldsymbol{\psi}(t) - \mathbf{Z}\mathbf{B}\boldsymbol{\psi}(t))'\}dt \\
&= \int_T \{(\mathbf{D} - \mathbf{Z}\mathbf{B})\boldsymbol{\psi}(t)\boldsymbol{\psi}'(t)(\mathbf{D} - \mathbf{Z}\mathbf{B})'\}dt \\
&= tr\{(\mathbf{D} - \mathbf{Z}\mathbf{B})\mathbf{J}_\psi(\mathbf{D} - \mathbf{Z}\mathbf{B})'\} \\
&= tr\{\mathbf{D}\mathbf{J}_\psi\mathbf{D}' - \mathbf{D}\mathbf{J}_\psi\mathbf{B}'\mathbf{Z}' - \mathbf{Z}\mathbf{B}\mathbf{J}_\psi\mathbf{D}' + \mathbf{Z}\mathbf{B}\mathbf{J}_\psi\mathbf{B}'\mathbf{Z}'\}
\end{aligned} \tag{5}
$$

where $\mathbf{J}_\psi = \int_T \psi(t)\psi(t)'$ is a matrix of basis function with $G \times G$ dimensions. we must solve for $\mathbf{B}$. Let the first derivative with respect to $\mathbf{B}$ is zero. We obtain:

$$-2(\mathbf{Z}'\mathbf{D}\mathbf{J}_\psi) + 2(\mathbf{Z}'\mathbf{Z}\mathbf{B}\mathbf{J}_\psi) = 0$$

$$\rightarrow \quad \mathbf{Z}'\mathbf{D}\mathbf{J}_\psi = \mathbf{Z}'\mathbf{Z}\mathbf{B}\mathbf{J}_\psi$$

By matrix vectorization, the solution of $\mathbf{B}$ is:

$$vec(\mathbf{Z}'\mathbf{D}\mathbf{J}_\psi) = vec(\mathbf{Z}'\mathbf{Z}\mathbf{B}\mathbf{J}_\psi)$$

$$\rightarrow \quad vec(\hat{B}) = (\mathbf{J}_\psi \otimes \mathbf{Z}'\mathbf{Z})^{-1} vec(\mathbf{Z}'\mathbf{D}\mathbf{J}_\psi)$$

where vec($\mathbf{B}$) is a column vector of length $(\sum_{k=1}^{K} M_k) \times G$.

3.3 Goodness of Fit [42]

One way extended from conventional linear model of accessing the fit of a functional leaner model is to consider the square correlation function[56]:

$$R^2(t) = 1 - \frac{\sum_i \{\hat{y}_i(t) - y_i(t)\}^2}{\sum_i \{y_i(t) - \bar{y}(t)\}^2}$$

R^2 measures the proportion of the total sample variance of the responses explained by the model [18]. For each individual function, an overall R^2 measurement defined by:

$$R^2(t) = 1 - \frac{\int \{\hat{y}_i(t) - y_i(t)\}^2 dt}{\int \{y_i(t) - \bar{y}(t)\}^2 dt}$$

Ramsay and Silverman(2005) [56] conceive of an F-ratio function for the fit:

$$\hat{y}_i(t) - \bar{y}(t) = \sum_{j=1}^{J_0} C_{ij} (\sum_{k=1}^{K_0} B_{jk} \psi_k(t)) = \sum_{j=1}^{J_0} C_{ij} \psi_j(t)$$

Here $K_0 - 1$ is the degrees of freedom to the point-wise sum of squares $\sum_i \{\hat{y}_i(t) - \bar{y}_i(t)\}^2$, $n - K_0$ is the degrees of freedom to the residual sum of squares $\sum_i \{y_i(t) - \hat{y}_i(t)\}^2$. So, F-ratio function can be constructed by:

$$FATIO(t) = \frac{\sum_i \{\hat{y}_i(t) - \bar{y}(t)\}^2 / (K_0 - 1)}{\sum_i \{y_i(t) - \hat{y}_i(t)\}^2 / (n - K_0)}$$

The parameter J_0 and K_O can be chosen by different methods, in which the choice of the appropriate method is probably subjectively. However, $\psi_j(t)$ might not be the best fit of $\hat{y}_i(t)$ to the true observed value, so the F-ratio could be used only as an approximation [56].

4 Applications: Case Study on Climate Data

In this section, we illustrate the insights one can obtain by using FDA techniques on historical climate data. This data set contains 4 years of high temporal resolution(hourly measurements) data of various weather attributes, temperature(K), humidity(%), air pressure(hPa), wind direction(meteorological degrees) and wind speed(m/s).

Original data are hourly weather time-series for 36 cities, 3 cities in Canada, 27 cities in USA, and the rest are Israel cities, which acquired using Weather API on the OpenWeatherMap Website. It is available under the ODbl License. The whole period of interest in this paper is 2013-01-01 01:00 to 2016-12-30 23:00.

Three samples extracted from the original data set will be used in this thesis. The first sample contains monthly records of all variables during the years from 2013 to 2016. For each city, the time span is aligned such that month 1 is the first month of the first year(2013). Furthermore, these monthly records by aggregating observed hourly data within each month.

Sample two is the records measured in weeks of all variables during the years 2013-2016. For each city, the time domain is aligned such that week 1 is the first week of the first year(2013). Weekly records calculated by aggregating observed hourly data within each week.

The third sample is the daily records of all variables during the years 2013-2016. Same as samples 1 and 2, daily records obtained by aggregating observed hourly data within a day. Individuals with any NA value were excluded from original data set before the aggregation. In this thesis, three samples will be tested, but we are concentrating on sample two measured by week when dealing with modelling.

4.1 Variables

Our observed data have five variables, they are:

Temperature: Temperature were measured in Kelvin scale.

Humidity: Humidity is a natural part of the our atmosphere, it comes from the amount of water vapor in the air.

Pressure: Pressure is the pressure within the atmosphere of Earth. The standard atmosphere (symbol: atm) is a unit of pressure defined as 101325 Pa (1013.25 hPa; 1013.25 mbar). Data in our case recorded by hPa unit.

Wind direction: Wind direction is reported by the direction from which it originates. Wind direction were reported in degrees in our case.

Wind speed: Weather forecasts typically give the direction of the wind along with its speed. For instance, a "southerly wind at 4m/s" is a wind blowing from the north at a speed of 4m/h.

A brief explanation of these variables is presented, along with a short summary of any

manipulations that has been made. Table 1 and 2, together with tables 6, 7 and 8 in Appendix, includes descriptive statistics for the variables used.

Table 1. Summary statistics for temperature(K) of the second sample(weekly records)

City	Mean	SD	Median	City	Mean	SD	Median
Vancouver	283.4067	6.070311	293.1192	Portland	284.2709	6.332019	294.3909
San.Francisco	287.2442	3.923411	293.6135	Seattle	283.69	5.815167	293.2045
Los.Angeles	290.3196	4.101892	299.8259	San.Diego	289.4416	3.594813	297.7546
Las.Vegas	292.6552	9.719126	310.2823	Phoenix	295.3331	8.88462	308.9593
Albuquerque	285.0698	8.654508	299.1572	Denver	282.2563	9.234211	296.2498
San.Antonio	293.1936	6.911027	303.9162	Dallas	292.1694	8.553323	306.086
Houston	293.7134	6.902858	303.2655	Kansas.City	286.5441	10.000147	300.7083
Minneapolis	279.612	12.585557	299.6084	Saint.Louis	286.2163	9.642247	301.5566
Chicago	283.1726	10.10729	300.5532	Nashville	288.1361	8.129465	300.4531
Indianapolis	284.4298	10.155913	300.5977	Atlanta	288.885	7.035304	298.3703
Detroit	282.7327	9.926664	299.9717	Jacksonville	293.8598	5.02865	301.4926
Charlotte	288.2368	7.439369	298.4307	Miami	297.9744	2.794086	301.9023
Pittsburgh	283.4743	9.30521	298.8626	Toronto	281.9634	9.398665	299.655
Philadelphia	285.0705	8.845086	301.3468	New.York	285.3736	9.099097	302.8306
Montreal	280.1779	11.316159	298.6286	Boston	283.8559	9.037263	299.7612
Beersheba	291.4523	5.796979	298.8389	Tel.Aviv.District	294.2746	5.105004	301.4063
Eilat	296.2143	6.127705	304.33	Haifa	293.479	4.166312	299.2359
Nahariyya	293.3869	4.451553	299.72	Jerusalem	291.0125	5.381442	298.3205

Table 2. Summary statistics for humidity(%) of the second sample(weekly records)

City	Mean	SD	Median	City	Mean	SD	Median
Vancouver	81.34106	9.084305	95.14894	Portland	76.12011	10.374347	91.45349
San.Francisco	75.77783	10.461149	100	Seattle	73.21957	10.773027	92.78571
Los.Angeles	59.68254	14.80109	90	San.Diego	66.87151	12.797352	89
Las.Vegas	29.86178	14.141567	63.33929	Phoenix	35.11888	12.763552	64.81395
Albuquerque	43.08925	15.091454	77.24224	Denver	55.54585	13.943318	86.20497
San.Antonio	68.79609	10.707267	93	Dallas	65.11062	9.348952	81
Houston	73.96538	9.585483	93	Kansas.City	68.61697	10.069572	93
Minneapolis	71.99296	8.78817	89.69512	Saint.Louis	70.81481	8.913923	93
Chicago	75.8878	9.973895	96.33333	Nashville	69.08622	12.354477	95.4
Indianapolis	71.31427	7.236865	88.15385	Atlanta	72.52109	12.025483	93
Detroit	75.00767	8.536407	93.42262	Jacksonville	81.56583	9.262791	100
Charlotte	74.04162	11.997385	93.58683	Miami	78.57541	8.253615	95
Pittsburgh	73.94956	9.317893	90.89706	Toronto	77.44293	7.90227	96.8869
Philadelphia	68.57433	9.345597	90	New.York	67.69277	11.038487	88.86667
Montreal	73.71401	8.22532	86.625	Boston	75.04555	8.324099	100
Beersheba	62.79452	15.445961	83	Tel.Aviv.District	64.70785	10.628055	86.18605
Eilat	45.66191	14.780239	87.14881	Haifa	79.91217	16.366038	100
Nahariyya	79.66081	18.430461	100	Jerusalem	70.24683	15.759945	93.02439

4.2 Data Preparation

We have 36 cities and some cities are landlocked while some cities are coastal. for example, Denver is a landlocked city, Boston is adjacent to the North Atlantic Ocean, and San Francisco is adjacent to the North Pacific Ocean. In Canada, Vancouver is in the North Pacific coastal city and Montreal and Toronto belong to North Atlantic coastal cities. Those cities in Israel are close to the Mediterranean Sea. As large bodies of water act as natural "heat reservoirs", we suppose that the different regions have distinct weather patterns. For all cities, we obtained hourly weather data onto 4 years. This allows us to look at changes across weeks or months, even years. From raw data, Figure 72 (see Appendix) presents observations from 2013-01-01 01:00 to 2013-12-31 23:00 measured by weeks.

In many cases, like in agriculture, the region is considered coastal if the coast directly impacts the region's weather, i.e. less than 7-10 km from the coastline. Based on that,

we separate these 36 cities into four groups, they are North Atlantic coastal cities, North Pacific Coastal cities, Mediterranean sea cities, and landlocked cities(see Table **??**). In order to intuitively have a look at that different regions have different weather patterns, we will use different colors to represent the corresponding geographic climates of the cities.

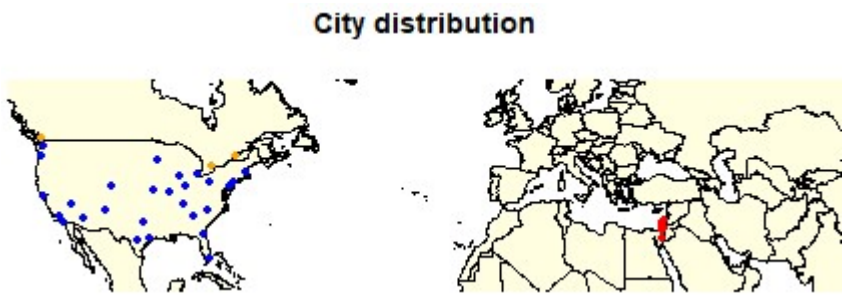


Figure 4. All cities location

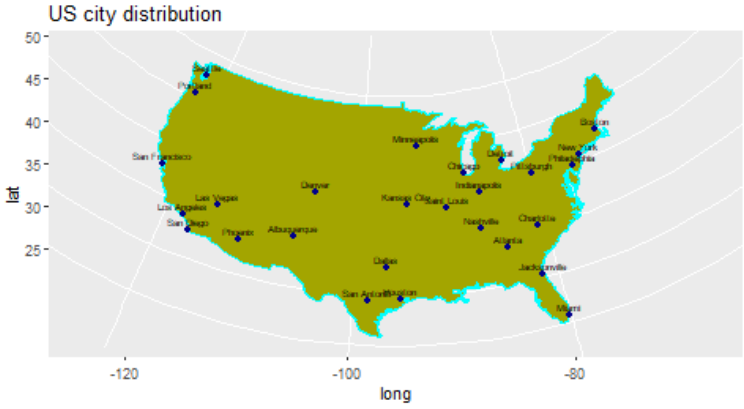


Figure 5. All cities location

Figure 6. Canada cities

Figure 7. Israel cities

In figure 72(see Appendix), we concentrate on the sample 2 giving the weekly records of different climate indicators over one year period. These are the individuals for which we have full information over this period. Each data point, represents the observed value of temperature, humidity, air pressure, wind direction, wind speed of every city.

Table 3. City Groups

Region	**City List**
Atlantic coastal	Montreal, Toronto, Houston, Jacksonville, Miami, Philadelphia, New York and Boston
Pacific coastal	Vancouver, Portland, San Francisco, Seattle, Los Angeles, San.Diego and Las.Vegas
Continental	Las.Vegas, Phoenix, Albuquerque, Denver, San Antonio, Dallas, Kansas.City, Minneapolis, Saint Louis, Chicago, Nashville, Indianapolis, Atlanta, Detroit, Charlotte and Pittsburgh
Mediterranean coastal	Beersheba, Tel.Aviv.District, Eilat, Haifa, Nahariyya, Jerusalem

An immediate indication is the variability for each individual over time. Fig 72 shows the raw data for each city. Clear that the temperature rise in the summer, fall in the winter. We expect that different regions will have different weather patterns. In most cases, continental cities show high peakedness which is cold winters and hot summers and coastal cities display little amplitude which is cool winters and summers. Except that, we still found some difference in level, for example, for those cities on the Pacific coast will have similar weather, but the temperature of city Las Vegas in the middle year is much higher than the other Pacific coastal cities. Similarly, in winter, city Jacksonville and Huston have a higher temperature than other Atlantic coastal cities. The resource of humidity in the atmosphere is water mass on earth. Relative humidity measures the actual amount of moisture in the air as a percentage of the maximum amount of moisture the air can hold. For the continental cities, the summer is the least humid months and

winter is the most humid. For the coastal cities, the humidity variation between winter and summer is lower than in continental cities, as we see the change range in coastal cities is smaller than in other regions. Comparing the curves representing the annual variations of pressure, we notice one Mediterranean city reaches far lowest from Mar to May, which is quite different from other cities, so does one pacific ocean city.

4.3 Representing Functional Data

In this section, we construct for each individual function of time that represents the curves of variables. Figure 77 provides examples of some curves (functions). Smoothing techniques were performed using B-spline basis of order 4 and Fourier basis of 5 nbasis. B-spline of order 4 is known as the cubic B-spline, which is a popular choice that providing a more flexible function. The ability of B-splines and Fourier bases to approximate smooth curves observed is tested on sample 2 (See Figure 77, 78 in Appendix). The first step was to choose the smoothing parameter λ. The smooth with roughness penalties were performed with respect to the second derivatives for B-spline basis and harmonic acceleration operation was used for Fourier basis. The penalty parameter λ was set by using GCV criterion. Results are shown in Table 4.

Table 4. Best lambda value and the degree of freedom for sample 2

Functional Variables	Type of Basis functions	K	**Best λ**	Minimum **GCV**	**df**
Temperature	Fourier	5	10000	208.3558	3.943231
	B-splines	4	100	211.5421	6.802356
Humidity	Fourier	5	10000	3462.731	3.894412
	B-splines	4	100	3320.415	6.802356
Pressure	Fourier Basis	5	1000	2842.8	4.779947
	B-Splines	4	10	2583.211	11.282282
Wind direction	Fourier	5	1000	47206.31	4.779947
	B-splines	4	10	35382.59	11.282282
Wind speed	Fourier	5	10000	18.54585	3.894412
	B-splines	4	100	16.50470	6.802356

4.3.1 Temperature

For weekly temperature data smoothing by Fourier basis, figure 8 indicated an optimal value of smoothing at $\lambda = 10^4$ because minimum GCV value obtained. At that value, $df(\lambda) = 3.943231$.

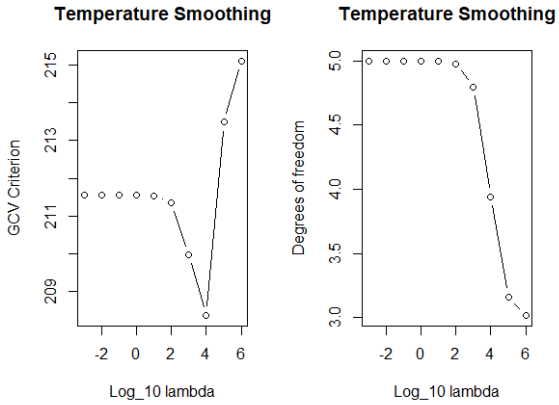


Figure 8. GCV for lambda values $\log_{10}(\lambda) = -2, 0, 2, 4, 6$, for weekly temperature records. The roughness penalty was defined by harmonic acceleration

Compare the smoothing mean curves to the observed data, the results are shown in Figure 9, the raw data are apparently non-smooth (see Figure 72). However, it is reasonable to assume that there is an underlying smoothly varying process that encapsulates the dynamics behind observed data. The deviations from a smooth function can be treated as aberrations or as instrumental errors. To get a sense of how well these curves explain the data, we have plot the residuals for temperature data along with the Fourier basis, the smallest standard deviation of the residuals from which is estimated is Haifa with 1.1K, and Kansas with 3.1K for the worst fitted curves (See Figure 9 (iv)). The residuals along with the B-spline basis fit is also provided (See Figure 9 (iii)-(v)), the smallest standard deviation of the variation of the actual temperature around the curves smoothing by B-spline basis is Haifa with 1.1K and biggest is Kansas with 2.9K.

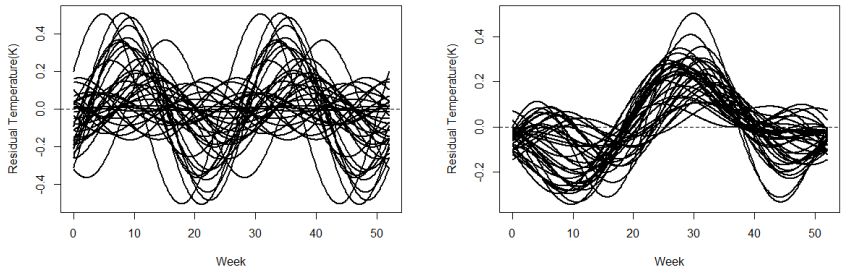


Figure 10. The smoothed residual functions for the temperature data. Figure in left side represents the functions smoothed by Fourier basis; figure in right side represents the functions smoothed by B-spline basis.

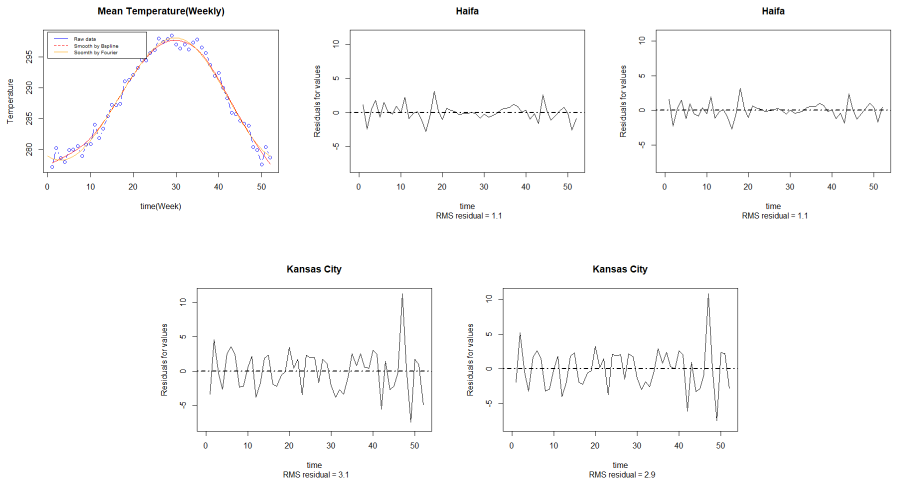


Figure 9. (i) Fitted mean curve of temperature. Mean function smoothing by B-spline and Fourier basis presented. Blue line indicates the weekly mean value of temperature. Red solid line represents the mean curves smooth by B-spline basis and orange solid line represents the mean curves smooth by Fourier Basis. (ii) Best fitted temperature curve smoothing by Fourier basis using GCV. (iii) Best fitted curve smoothing by B-spline basis using GCV. (iv) Worse fitted temperature curve smoothing by Fourier basis. (v) Worse fitted temperature curve smoothing by B-spline basis.

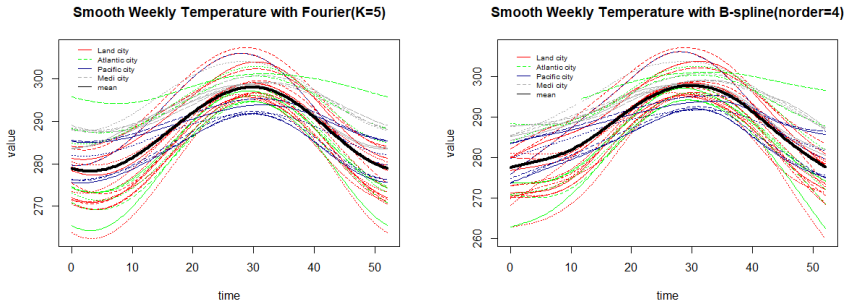


Figure 11. The smoothed functions for the temperature data. Figure in left side represents the functions smoothed by Fourier basis; figure in right side represents the functions smoothed by B-spline basis.

As can be seen in Figure 12, the temperature variance surface is the direct representation of the temperature over time, which shows the mean temperature curves as well as the covariance of the average weekly temperature of all cities. Looking at the smooth temperature curve (Figure 11), it is evident that the temperature curves of Mediterranean

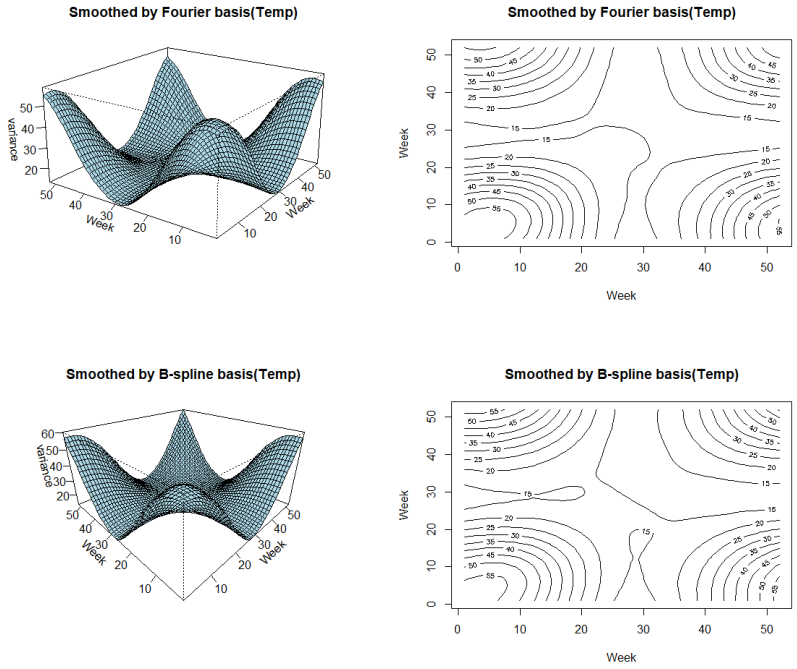


Figure 12. (i) Covariance function value using Fourier basis expansion. (ii) Same surface by contour plotting of Covariance function using Fourier basis expansion. (iii) Convariance function value using B-spline basis expansion. (iv) Contour plotting of covariance function using B-spline basis expansion.

coastal cities exhibit much higher weather patterns than the mean temperature (Figure 11(i) - (ii)). During winter time, the weekly temperature in Pacific coastal cities appear marginally higher than mean temperature, while in summer, the temperature in coastal is higher than the mean temperature. These patterns actively point that the yearly variation of average temperatures is higher in continental cities than in coastal cities.

4.3.2 Humidity

Figure 13 (i) depicts the fitted mean curves given on the observed mean humidity applying the information from Table 4. In this case, Fig.13 (ii) - (iii) are two plots of the residuals for humidity data. Middle plot plotted along with the Fourier basis, the standard deviation of the residuals of best fitted curves from which is estimated is Miami with 6 units. The residuals along with the B-spline basis fit is also provided in right plot, the standard deviation of best fitted curve is Miami with 6 units. Total residual functions of humidity shown on the Figure 14.

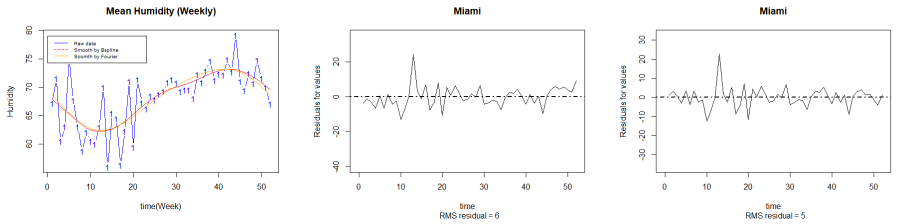


Figure 13. (i) Fitted mean curve of humidity. Mean function smoothing by B-spline and Fourier basis presented. Blue line indicates the weekly mean value of humidity. Red solid line represents the mean curves smooth by B-spline basis and orange solid line represents the mean curves of humidity smoothing by Fourier Basis. (ii) Residual plot of best fitted humidity curves smoothing by Fourier basis using GCV. (iii) Residual plot of best fitted humidity curve smoothing by B-spline basis using GCV.

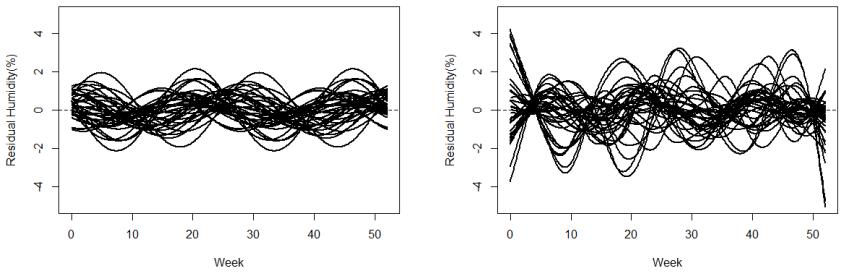


Figure 14. (i) The smoothed residuals functions for humidity using Fourier basis. (ii) The smoothed residuals functions for humidity using B-spline basis.

4.3.3 Pressure

Applying the information in Table 4 to the observed weekly average pressure, the fitted mean curves shown in Figure 15 (i). In the case of pressure, Figure 15 (ii) - (iii) are two plots of the residuals for pressure data. Middle plot plotted along with the Fourier basis, the standard deviation of the residuals of best fitted curves from which is estimated is Haifa with 3 hPa. The residuals along with the B-spline basis fit is also provided in right plot, the standard deviation of best fitted curve is Haifa with 2.4 hPa. Total residual functions of pressure shown on the Figure 16.

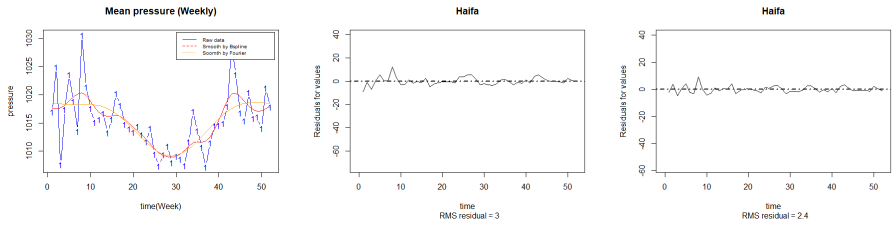


Figure 15. (i) Fitted mean curve of pressure. Mean function smoothing by B-spline and Fourier basis presented. Blue line indicates the weekly mean value of pressure. Red solid line represents the mean curves smooth by B-spline basis and orange solid line represents the mean curves smooth by Fourier Basis. (ii) Residual plot of best fitting curves of pressure smoothing by Fourier basis using GCV. (iii) Residual plot of best fitting curves of pressure smoothing by B-spline basis using GCV.

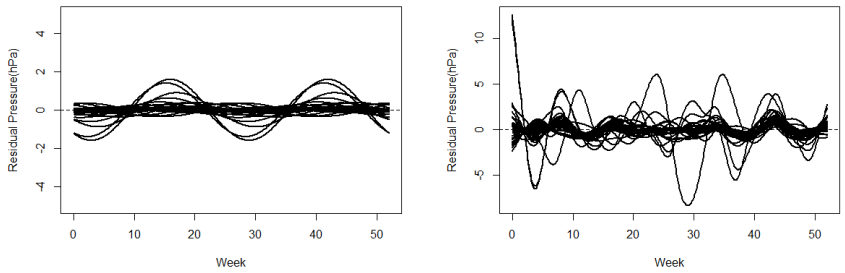


Figure 16. (i) The smoothed residuals functions for pressure using Fourier basis. (ii) The smoothed residuals functions for pressure using B-spline basis.

4.3.4 Wind Direction

Figure 17 (i) depicts the fitted mean curves of wind direction produced on the observed mean value using the information from Table 4. Figure 17 (ii) - (iii) are two plots of the residuals for wind direction. Middle plot plotted along with the Fourier basis, the standard deviation of the residuals of best fitted curves from which is estimated is Tel Avlv District with 21.3 units. The residuals along with the B-spline basis fit is also provided in right plot, the standard deviation of best fitted curve is same city with 16.7 units. Total residual functions of wind speed shown on the Figure 18.

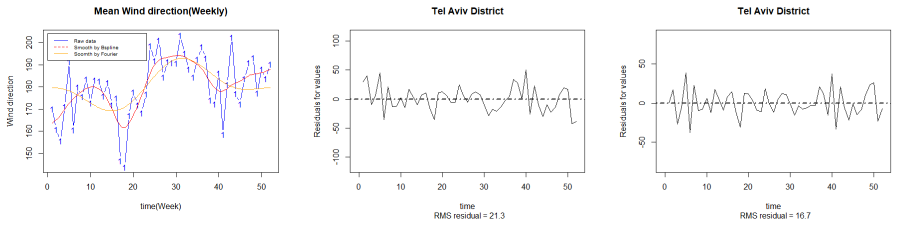


Figure 17. (i) Fitted mean curve of wind direction. Mean function smoothing by B-spline and Fourier basis presented. Blue line indicates the weekly mean value of wind direction. Red solid line represents the mean curves smooth by B-spline basis and orange solid line represents the mean curves smooth by Fourier Basis. (ii) Residual plot of best fitting curves of wind direction smoothing by Fourier basis. (iii) Residual plot of best fitting curves of wind direction smoothing by B-spline basis.

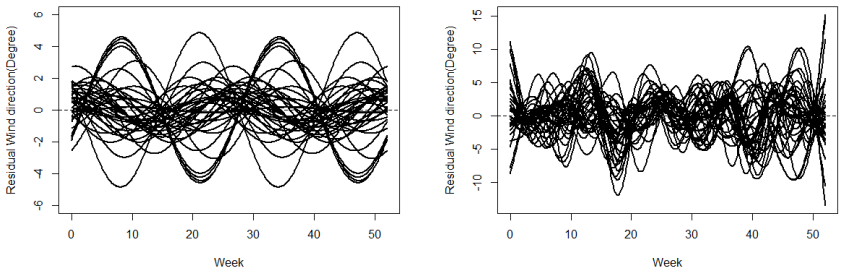


Figure 18. (i) The smoothed residuals functions for wind direction using Fourier basis. (ii) The smoothed residuals functions for wind direction using B-spline basis.

4.3.5 Wind Speed

Figure 19 (i) shows the fitted mean curves of wind speed produced from the observed mean value utilizing the information from Table 4. Figure 19 (ii) - (iii) are two plots of the residuals for wind speed. Middle plot plotted along with the Fourier basis, the standard deviation of the residuals of best fitted curves from which is estimated is Los Angeles with 0.187 degree. The residuals along with the B-spline basis fit is also provided in right plot, the standard deviation of best fitted curve is Los Angeles with 0.169 degree. Total residual functions of wind direction shown on the Figure 20.

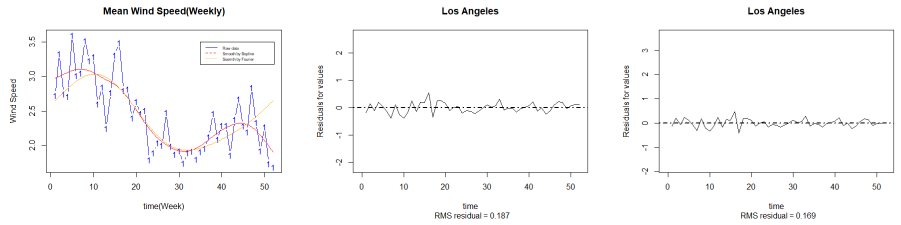


Figure 19. (i) Fitted mean curve of wind speed. Mean function smoothing by B-spline and Fourier basis presented. Blue line indicates the weekly mean value of wind speed. Red solid line represents the mean curves smooth by B-spline basis and orange solid line represents the mean curves smooth by Fourier Basis. (ii) Residual plot of best fitting curves of wind speed smoothing by Fourier basis using GCV. (iii) Residual plot of best fitting curves of wind speed smoothing by B-spline basis using GCV.

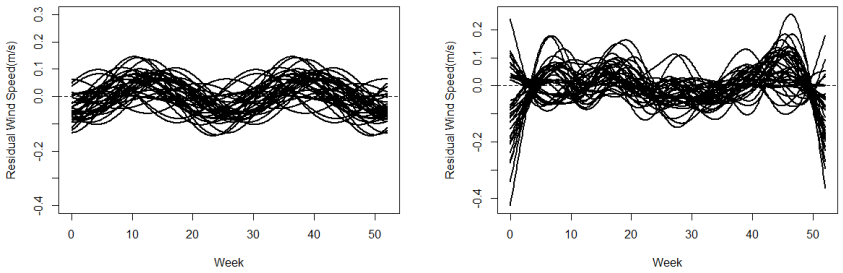


Figure 20. (i) The smoothed residuals functions for wind speed using Fourier basis. (ii) The smoothed residuals functions for wind speed using B-spline basis.

4.4 Functional Principal Component Analysis

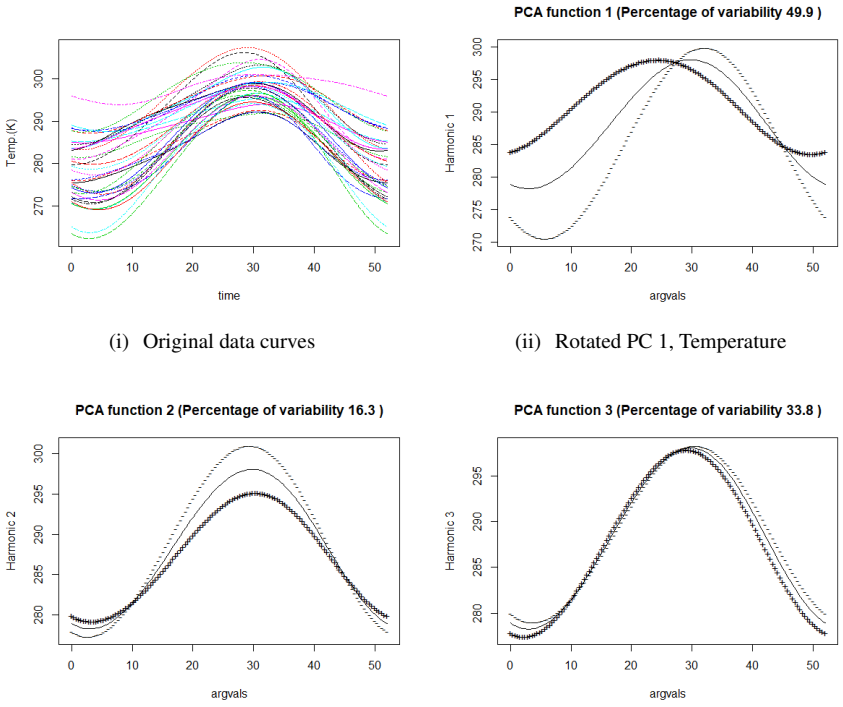
We explored the functional principal component analysis (fPCA) after smoothing curves using Fourier basis with 5 terms. In classical statistic, PCA seeks to reduce the dimensionality of a multivariate data into independent linear combinations of the variables. In FDA, the aim of fPCA is to reduce the dimensionality by finding those curves which capture the fundamental modes of variation of the data.

4.4.1 Temperature

As the original PCs mainly captured variations coming from the initial functions, we rotated original PCs using the **VARIMAX** rotation method. The fPCA of the temperature curves reveals that the first three principal components account for more than 95% of their variation about the mean curve. After **VARIMAX** rotation of these components, we

get the three components displayed in Figure 21 ((i)-(iv)), and the effects of adding and subtracting each rotated PCs from the mean function can be seen.

The first rotated component portrays primarily variations in Summer and Autumn and the second captures stronger variation in winter. The third components very much like the mean.



(i) Original data curves

(ii) Rotated PC 1, Temperature

(iii) Rotated PC 2, Temperature

(iv) Rotated PC 3, Temperature

Figure 21. The means plotted with the effect of each PC for weekly temperature data. The dashed "+" shows the effect of adding the PC to the mean, the dashed "-" shows the effect of subtracting the PC. The effects of the first three varimax-rotated components for the temperature data. The Fraction-of-Variance-Explained for each component is shown in the title.

In order to see if functional cluster or other features can be found, principal component scores for pairs of harmonics, as has been done in Figure 22 for temperature data. Most cities are falling into three clusters. The upper middle with the Pacific coastal, the upper left with Atlantic coastal and continental cities and lower left with the Mediterranean coastal cities. As stated in Chapter 2, the functional bagplot is considered as a mapping of the bivariate bagplot of the first two robust principal component scores to the functional curves[50]. Ordered by Tukey's halfspace location depth, the bivariate and functional bagplot of the temperature data displayed in Figure 22. The detected outliers in the temperature data is Minneapolis. Figure 24 display the bivariate and functional HDR

boxplot of the temperature data, the detected outliers is Minneapolis and Miami.

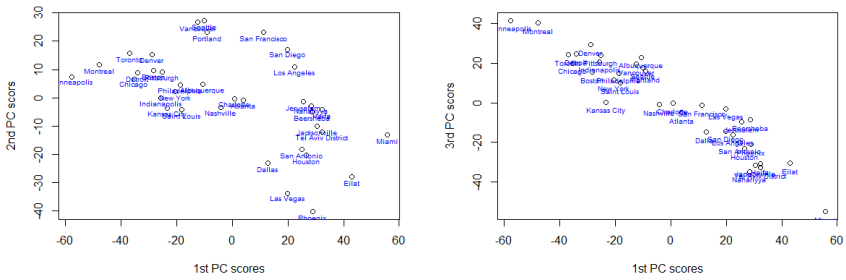


Figure 22. Scatter plots of the rotated principal component scores of the weekly temperature data. Selected stations are labeled.

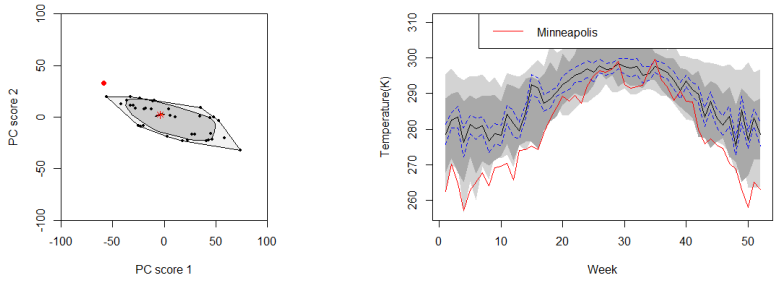


Figure 23. The bivariate bagplot and functional bagplot for temperature data is provided. The bivariate principal component scores can be ordered by Tukey's halfspace location depth and plotted in a two-dimensional graph.

4.4.2 Humidity

For fPCA of humidity, the first three principal components originally account for more than 95% of their variation about the mean curve. After **VARIMAX** rotation, three rotated principal components displayed in Figure 25 ((ii) - (iv)). The first rotated PC1 portrays primarily variations in Dec till Mar, the rotated PC2 captures stronger variation from Apr till Aug. Rotated PC3 captures more variation from Aug till Dec. The principal component scores for pairs of harmonics shown on Figure 26. Bagplot of humidity is provided in Figure 27. There are five outliers detected, they are Los Angeles, Jacksonville, Las Vegas, Nahariyya and Phoenix. Figure 28 display the bivariate and functional HDR boxplot of the humidity data. The detected outliers is Los Angeles and Denver.

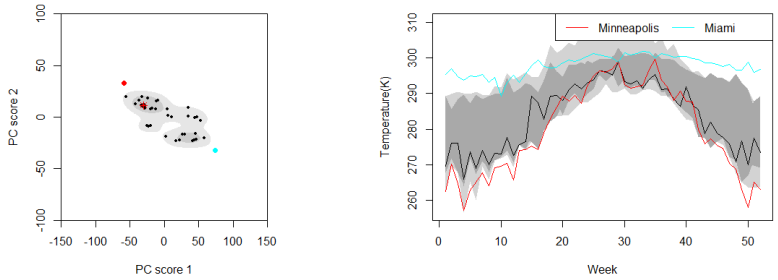
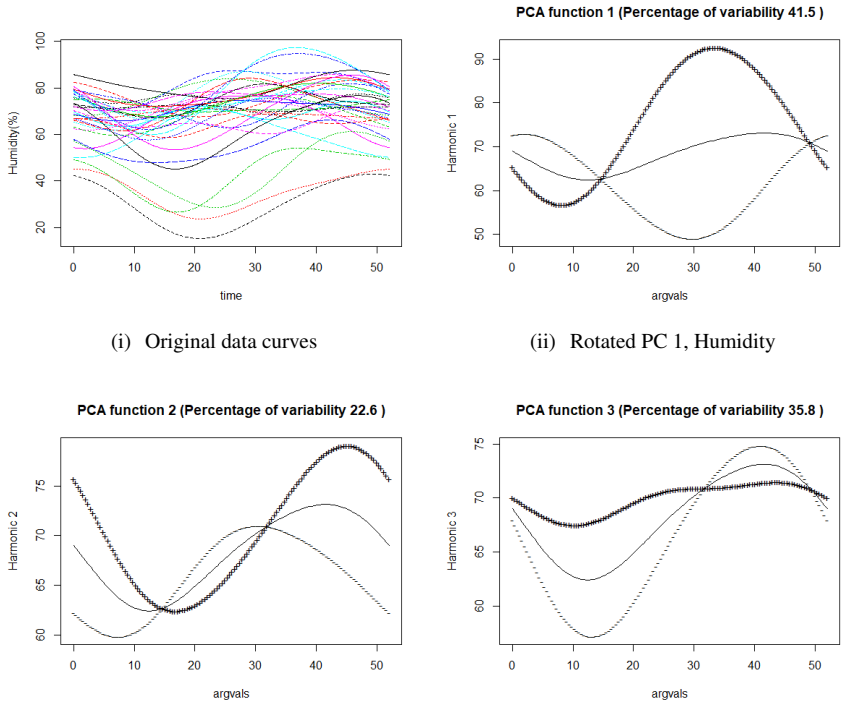


Figure 24. The bivariate boxplot and functional boxplot for temperature data is provided. The bivariate principal component scores can be ordered by the highest density regions and plotted in a familiar two-dimensional graph. The dark and light gray regions show the 50% HDR and the 95% outer HDR, respectively. These correspond directly to the equivalent regions with similar colors and shading in the functional HDR boxplot.



(i) Original data curves

(ii) Rotated PC 1, Humidity

(iii) Rotated PC 2, Humidity

(iv) Rotated PC 3, Humidity

Figure 25. The means plotted with the effect of each PC for weekly humidity data. The dashed "+" shows the effect of adding the PC to the mean, the dashed "-" shows the effect of subtracting the PC. The effects of the first three varimax-rotated components for the humidity data. The Fraction-of-Variance-Explained for each component is shown in the title.

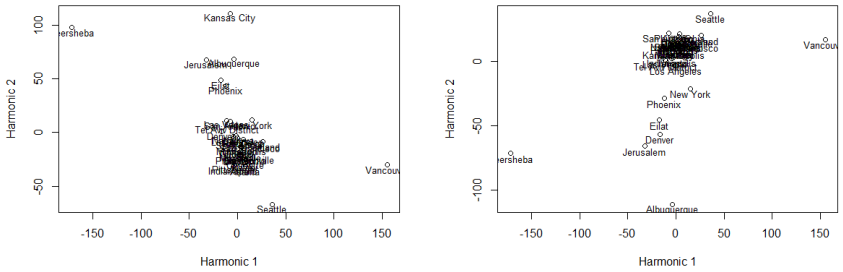


Figure 26. Scatter plots of the rotated principal component scores of the weekly humidity data. Selected stations are labeled.

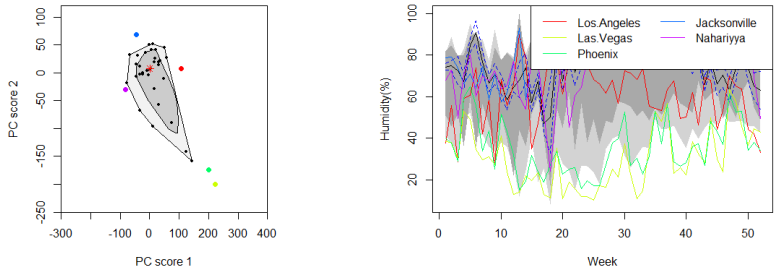


Figure 27. The bivariate bagplot and functional bagplot for humidity data is provided. The bivariate principal component scores can be ordered by Tukey's halfspace location depth and plotted as a two-dimensional graph.

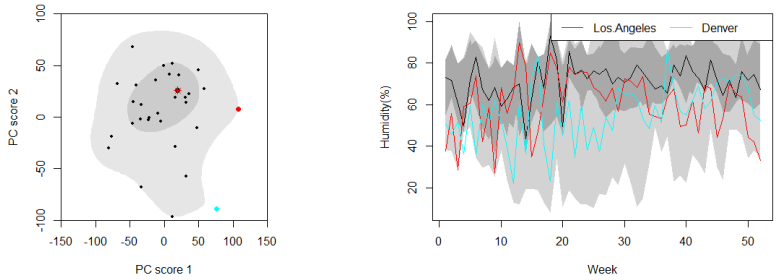
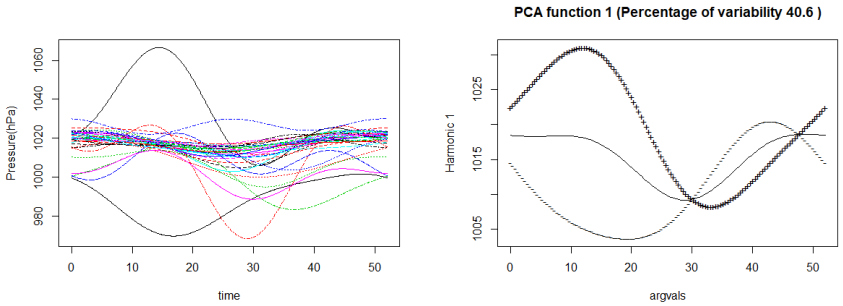


Figure 28. The bivariate boxplot and functional boxplot for humidity data is provided. The bivariate principal component scores can be ordered by the highest density regions and plotted in a two-dimensional graph. The dark and light gray regions show the 50% HDR and the 95% outer HDR, respectively. These correspond directly to the equivalent regions with similar colors and shading in the functional HDR boxplot.

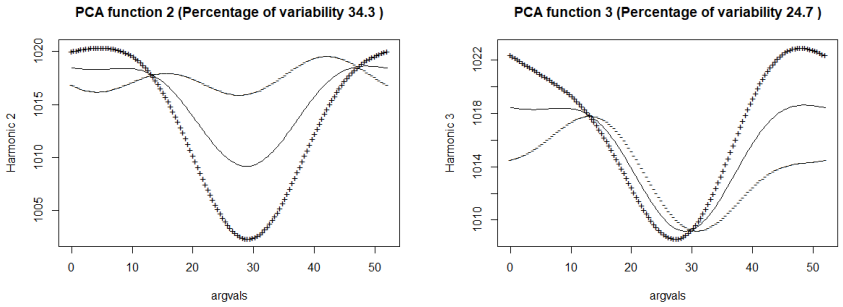
4.4.3 Pressure

After rotation, the first three rotated principal components displayed in Figure 29 ((ii)-(iv)). The first rotated component accounts for 40% of the total variance that portrays primarily variations from mid-July to Nov. The second component captures more variations from Dec to Apr, which accounts for nearly 35% of the total variance. The third component captures more variation from Apr till July accounting for 25% of the total variance. The principal component scores for pairs of harmonics shown on Figure 26. Tukey's halfspace bagplot of pressure is provided in Figure 27. Two curves detected as outliers, they are Kansas and Beersheba. Figure 32 display the bivariate and functional HDR boxplot of the pressure. Two functions were identified as outliers, Vancouver and Seattle.



(i) Original data curves

(ii) Rotated PC 1, Pressure



(iii) Rotated PC 2, Pressure

(iv) Rotated PC 3, Pressure

Figure 29. The means plotted with the effect of each PC for weekly pressure data. The dashed "+" shows the effect of adding the PC to the mean, the dashed "-" shows the effect of subtracting the PC. The effects of the first three varimax-rotated components for the pressure data. The Fraction-of-Variance-Explained for each component is shown in the title.

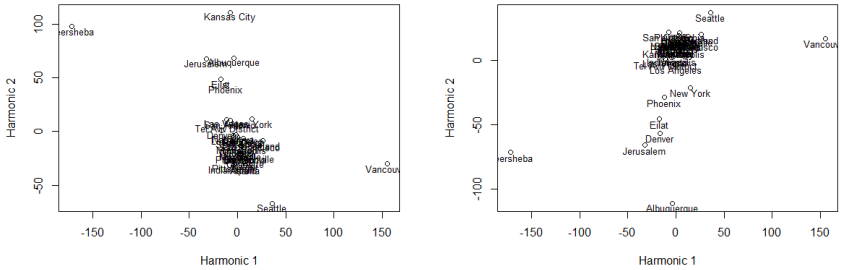


Figure 30. Scatter plots of the rotated principal component scores of the weekly pressure data. Selected stations are labeled.

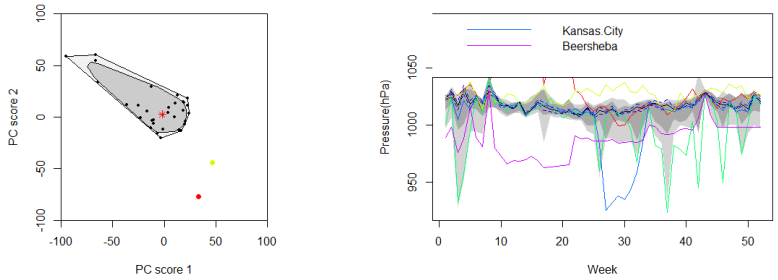


Figure 31. The bivariate bagplot and functional bagplot for pressure data is provided. The bivariate principal component scores can be ordered by Tukey's halfspace location depth and plotted in a two-dimensional graph.

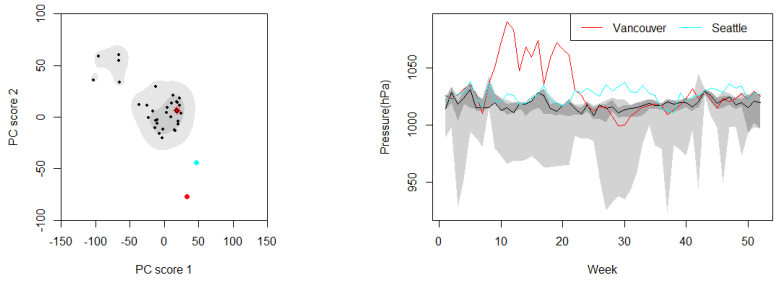
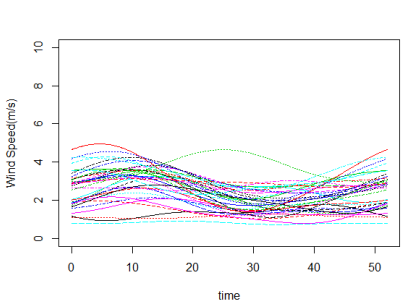


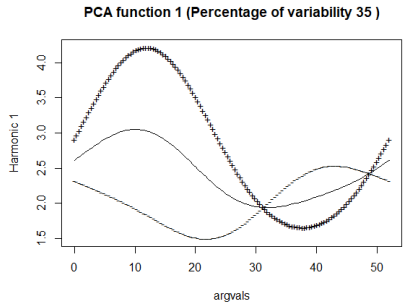
Figure 32. The bivariate boxplot and functional boxplot for pressure data is provided. The dark and light gray regions show the 50% HDR and the 95% outer HDR, respectively. These correspond directly to the equivalent regions with similar colors and shading in the functional HDR boxplot.

4.4.4 Wind Speed

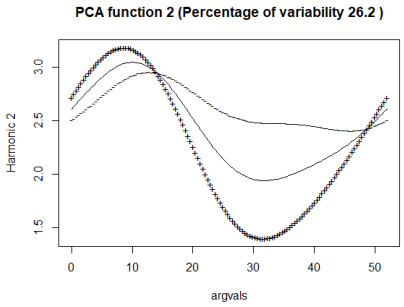
The three rotated principal components displayed in Figure 33 ((ii)-(iv)). The first rotated component accounts for 35% of the total variance that reveals primarily variations from mid-July to Nov. The second component captures more variations from Dec to Apr, which accounts for nearly 26.2% of the total variance. The third component captures stronger variation from Apr till mid-July accounting for 38.7% of the total variance. The principal component scores for pairs of harmonics shown on Figure 34. Tukey's halfspace bagplot is provided in Figure 35. Two curves detected as outliers, they are Toronto and Eilat. Figure 32 display the bivariate and functional HDR boxplot of the pressure. Los Angeles and Eilat were detected as outliers.



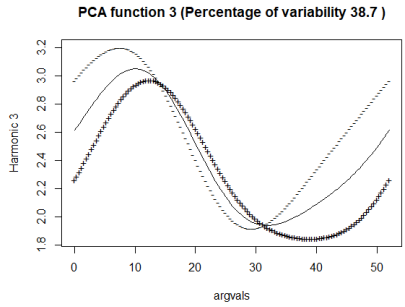
(i) Original data curves



(ii) Rotated PC 1, Wind Speed



(iii) Rotated PC 2, Wind Speed



(iv) Rotated PC 3, Wind Speed

Figure 33. The means plotted with the effect of each PC for weekly wind speed. The dashed "+" shows the effect of adding the PC to the mean, the dashed "-" shows the effect of subtracting the PC. The effects of the first three varimax-rotated components for the wind speed. The Fraction-of-Variance-Explained for each component is shown in the title.

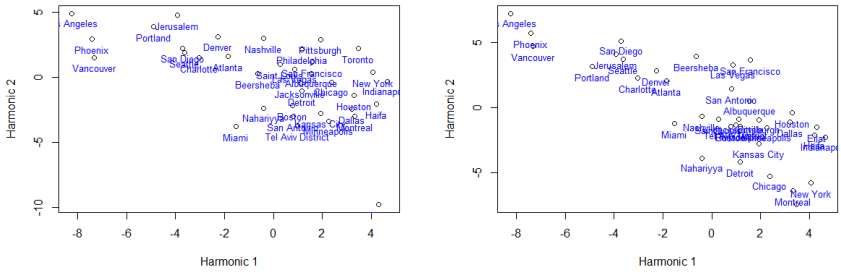


Figure 34. Scatter plots of the rotated principal component scores of the weekly wind speed data. Selected stations are labeled.

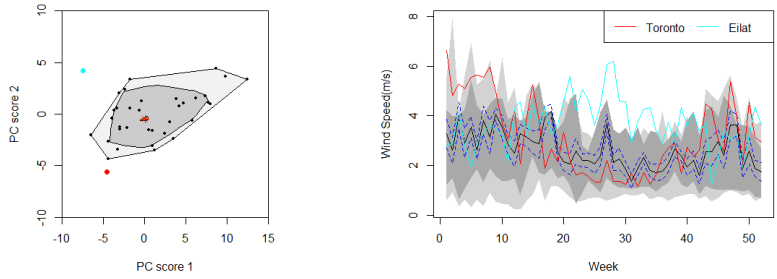


Figure 35. The bivariate bagplot and functional bagplot for wind speed data is provided. The bivariate principal component scores can be ordered by Tukey's halfspace location depth and plotted in a graph.

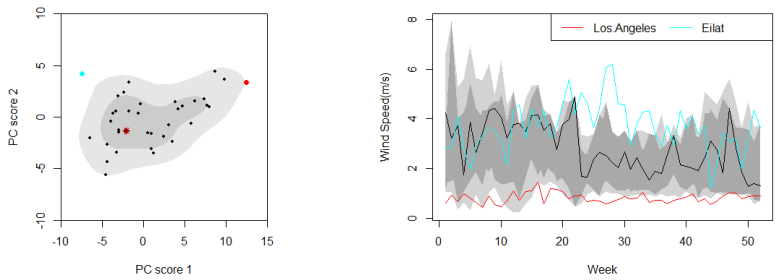
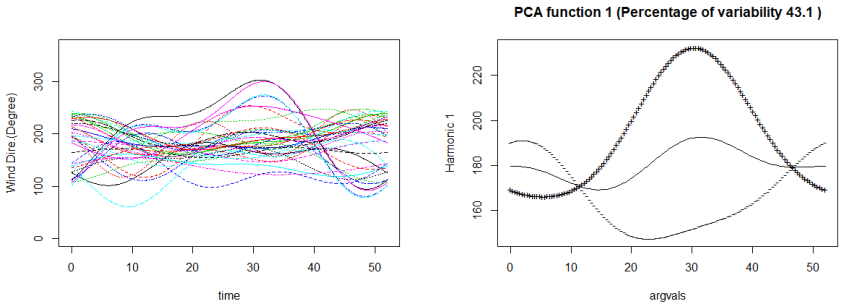


Figure 36. The bivariate boxplot and functional boxplot for wind speed data is provided. The dark and light gray regions show the 50% HDR and the 95% outer HDR, respectively. These correspond directly to the equivalent regions with similar colors and shading in the functional HDR boxplot.

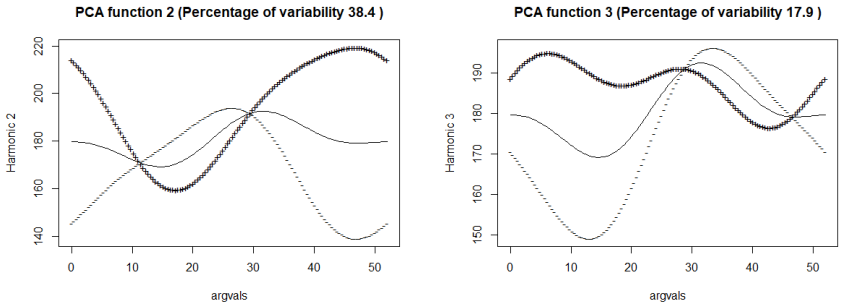
4.4.5 Wind Direction

After rotation, three rotated principal components displayed in Figure 37 ((ii)-(iv)). The first rotated component accounts for 43.1% of the total variance that reveals primarily variations from mid-Nov to Mar. The second component captures more variations from Mar to mid-July, which accounts for 38.4% of the total variance. The third component captures stronger variation from mid-July to Nove accounting for 17.9% of the total variance. The principal component scores for pairs of harmonics shown on Figure 38. Tukey's halfspace bagplot is provided in Figure 39 and no outliers being detected. But when ordered by highest density region, two curves identified as outliers. They are Denver and Tel Aviv District, results are shown in Figure 40.



(i) Original data curves

(ii) Rotated PC 1, Wind direction



(iii) Rotated PC 2, Wind direction

(iv) Rotated PC 3, Wind direction

Figure 37. The means plotted with the effect of each PC for weekly wind direction data. The dashed "+" shows the effect of adding the PC to the mean, the dashed "-" shows the effect of subtracting the PC. The effects of the first three varimax-rotated components for the wind direction data. The explained variance for each component is shown in the title.

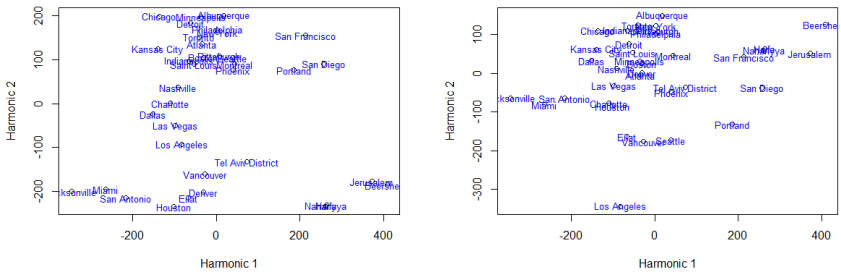


Figure 38. Scatter plots of the rotated principal component scores of the weekly wind direction data. Selected stations are labeled.

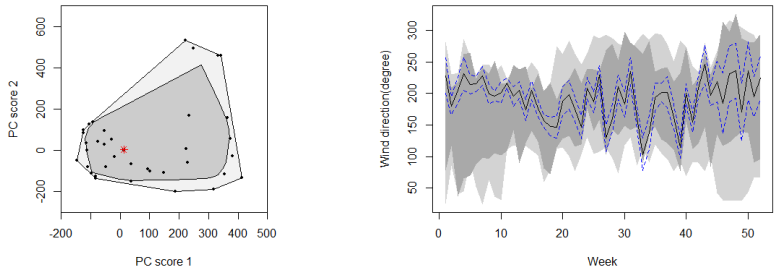


Figure 39. The bivariate bagplot and functional bagplot for wind direction is provided. The bivariate principal component scores can be ordered by Tukey's halfspace location depth and plotted in a two-dimensional graph.

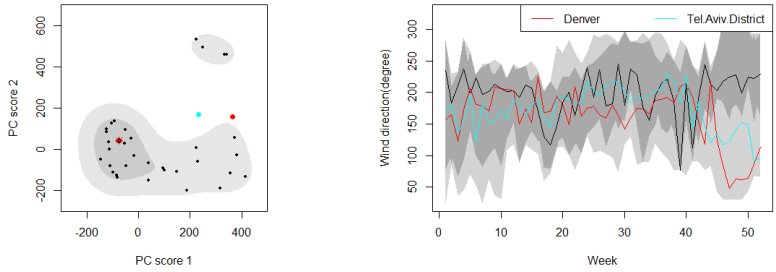


Figure 40. The bivariate boxplot and functional boxplot for wind direction data is provided. The bivariate principal component scores can be ordered by the highest density regions. The dark and light gray regions show the 50% HDR and the 95% outer HDR, respectively. These correspond directly to the equivalent regions with similar colors and shading in the functional HDR boxplot.

4.5 Functional Linear Model

4.5.1 Functional analysis of variance

In this section, we would describe existing tests for the one-way ANOVA problem for functional data, which is a natural way to quantity how much the pattern of annual variation of each climate indicator is attributed to different areas. In our case, we divided all city locations into four groups: Atlantic coastal cities, Pacific coastal cities, Continental cities and Mediterranean cities.

Consider $X_{i1}(t), X_{i2}(t), ..., X_{in_i}(t), i = 1, ..., k$, are k groups of independent random functions defined over a finite interval $T = [a, b]$. Assume that $X_{ij}(t), i = 1, 2, ..k, j = 1, ..., n_i$

are stochastic process with mean function $\mu_i(t), t \in T$ and its covariance function $\gamma(s,t), s,t \in T$, for $i = 1,...k$. Since we separated all cities into four groups, $k = 4$. In figure 11, it can be seen that the temperatures at the coastal cities are basically higher than those continental cities. Further investigate if the location statistically has an effect on the mean temperature curves of the Atlantic coastal cities, Pacific coastal cities, Continental cities and Mediterranean cities, is equivalent to the one-way ANOVA problem for functional data, This issue was addressed by Zhang(2013) [69]. One-way analysis of variance problem for functional data is to test the following null hypothesis

$$H_0 : \mu_1(t) = \mu_2(t) = \mu_3(t) = \mu_4(t), t \in T.$$

$$H_1 : \textit{The means are not all equal}.$$

R package **fdANOVA** [15] provides access to a wide range of overview of analysis of variance methods for functional data. In our case, we applied all the tests under considerations, getting the results given below. The p-values of the all results (Table 5) are less than the significance level 0.05, H_0 is rejected. Hence it can be concluded that location has an effect on the mean temperature curves of the Atlantic coastal cities, Pacific coastal cities, Continental cities and Mediterranean coastal cities. Same fANOVA process were tested on other variables, results are shown in Table (9, 10, 11 and 12 in Appendix). For monthly humidity data, p-value of most tests are bigger than 0.05 that the we don't have enough evidence to reject the null hypothesis. For weekly and daily humidity, the p-values of all tests are smaller than 0.05. Which means location has an effect on mean curves weekly humidity and daily humidity. For variable pressure, wind speed and wind direction, the p-value of all tests results are smaller than 0.05.

Table 5. Values of test statistics and p-values of all tests for monthly / weekly / daily temperature(K) data for [a,b]=[0,1]

Test	monthly temp.data		weekly temp.data		daily temp.data	
	Test stat.	p-value	Test stat.	p-value	Test stat.	p-value
CH	20510.98	0.0104	97531.51	0.0081	797934.9	0.0051
CS	20510.98	0.0049	97531.51	0.005	797934.9	0.0019
L2N	3846.623	0.002958811	18499.87	0.001503242	154817.7	0.0003885458
L2B	3846.623	0.001937893	18499.87	0.0009197085	154817.7	0.0001978849
L2b	3846.623	0.0065	18499.87	0.0051	154817.7	0.0017
FN	4.174943	0.007714193	4.243204	0.00451261	4.224115	0.001515482
FB	4.174943	0.0.007289066	4.243204	0.004035221	4.224115	0.00115749
Fb	4.174943	0.0233	4.243204	0.0183	4.224115	0.0114
GPF	4.319263	0.00187661	4.627195	0.0004611861	4.831433	4.042774e-05
Fmaxb	6.2333339	0.0094	9.589859	0.0022	21.38179	1e-04
TRP	-	1	-	1	-	1
FP	4.183326	0.008	4.252671	0.003	4.224533	0.001

Notes: TRP - tests based on K = 30 random projections and p-value ANOVA without permutation.

4.5.2 Analysis of Variance Model

Analysis of variance can be modelled as a linear model for functional responses with scalar covariates in the sense that variation in a functional response is decomposed into

functional effects through a scalar matrix. In our case, the effect of different regions on the shape of the climate variable curves is of interest.

For the i-th city in the k-th climate group, we have the model for temperature of following form:

$$Temp_{ik}(t) = \mu(t) + \alpha_{ik}(t) + \varepsilon_{ik}(t).$$

Here α_k represents the specific effects on temperature in climate group k, with a constraint $\sum_k \alpha_k(t) = 0$ for all $t \in T$.

Figure 41 displays the estimated regions effects, along with 95% pointwise confidence intervals estimated by the methods mentioned in the above methodology section. Figure 42 shows the composite effects $\mu + \alpha_k$. We see that the Atlantic coastal cities have a temperature about 5 Kelvin warmer than the mean temperature of all cities, but even more so in Winter than other seasons. The Pacific coastal cities are close to the mean temperature of all cities except the Winter and Spring. In Winter and Spring, the Pacific coastal cities are colder than mean temperature of all cities. The continental cities are cooler than average temperature of all cities, but close to mean temperature in Winter. In Summer, the Mediterranean cities are slightly warmer than mean temperature of all cities, but are colder in other seasons.

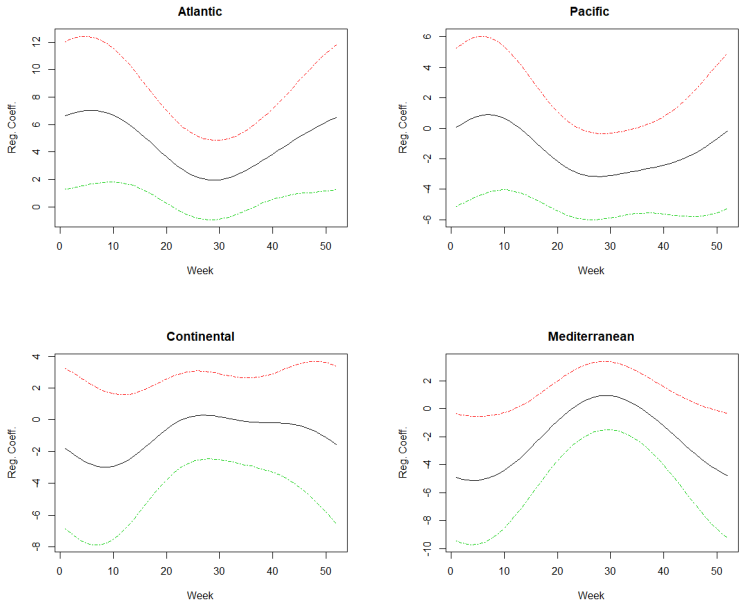


Figure 41. The region effects α_k for the temperature functions in the functional analysis of variance model. The effects $\alpha_k(t)$ are required to sum to 0 for all t. The dashed lines indicates 95% point-wise confidence intervals for the true effects.

We've also look at the squared correlation function (RSQ) and F-ratio functions (FRATIO). Results are shown in Figure 43. RSQ function considers the drop in error sum of squares

produced by taking climate region into effect relative to error sum of squares without using climate region information. The squared correlation is relatively high. F-ratio is everywhere higher than the 5% significant level. The difference between climate regions are substantially stronger in Summer and Autumn.

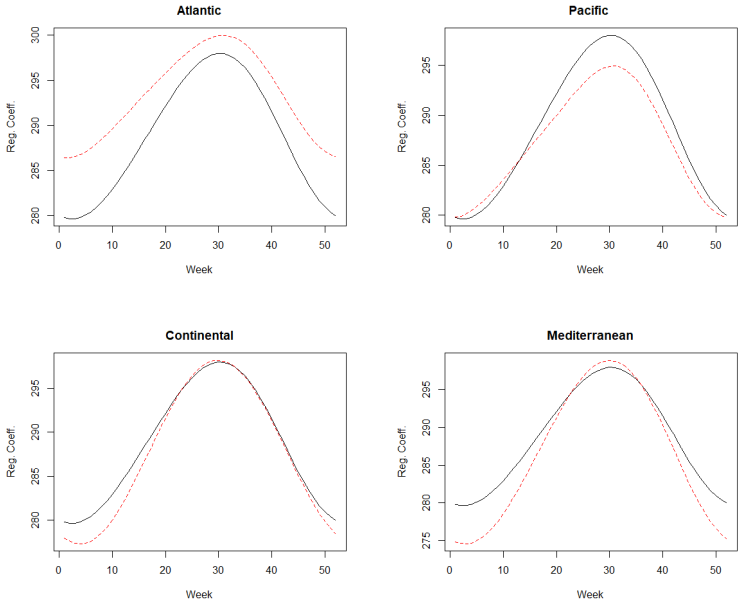


Figure 42. The dashed red curves are estimated climate zone temperature profile $\mu + \alpha_k$ for the temperature functions in the functional analysis of variance model. The solid black curve is the mean function μ of all cities.

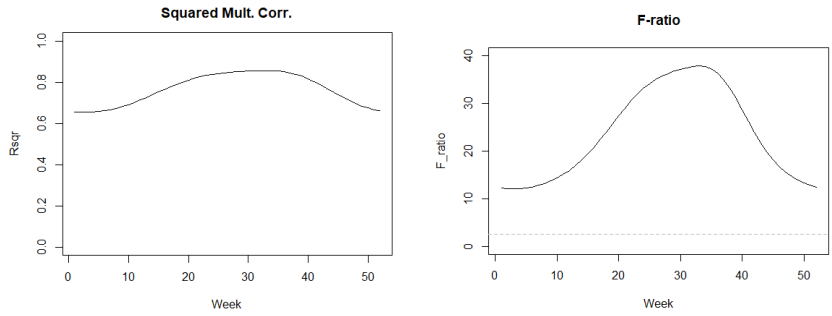


Figure 43. (i) Squared multiple correlation RSQ. (ii) The corresponding F-ratio function FRATIO. the dotted gray line indicates 5% significant level for the F-distribution with 5 and 32 degrees of freedom.

Same process are computed on other climate variables. For variable humidity, $Humi_{ik}$ is the humidity of i-th city in the k-th climate group, we have the model of following form:

$$Humi_{ik}(t) = \mu(t) + \alpha_{ik}(t) + \varepsilon_{ik}(t).$$

Here α_k represents the specific effects on humidity in climate group k, with a constraint $\sum_k \alpha_k(t) = 0$ for all $t \in T$.

Figure 44 displays the estimated regions effects, along with 95% pointwise confidence intervals estimated. Figure 45 shows the composite effects $\mu + \alpha_k$. We see that the Atlantic coastal cities are much lower humidity than the mean humidity of all cities in Spring the Summer, but are slightly in other season. But Pacific coastal cities are hold contrary situation. The Pacific coastal cities are higher humidity than mean level from April until Feb. The continental cities are much higher humidity than average humidity of all cities during whole year. For Mediterranean cities, the humidity is lower than the mean level, but even more so in Autumn and Winter.

Results of RSQ and FRATIO functions are shown in Figure 46. The squared correlation is relatively low. F-ratio is everywhere lower than the 5% significant level.

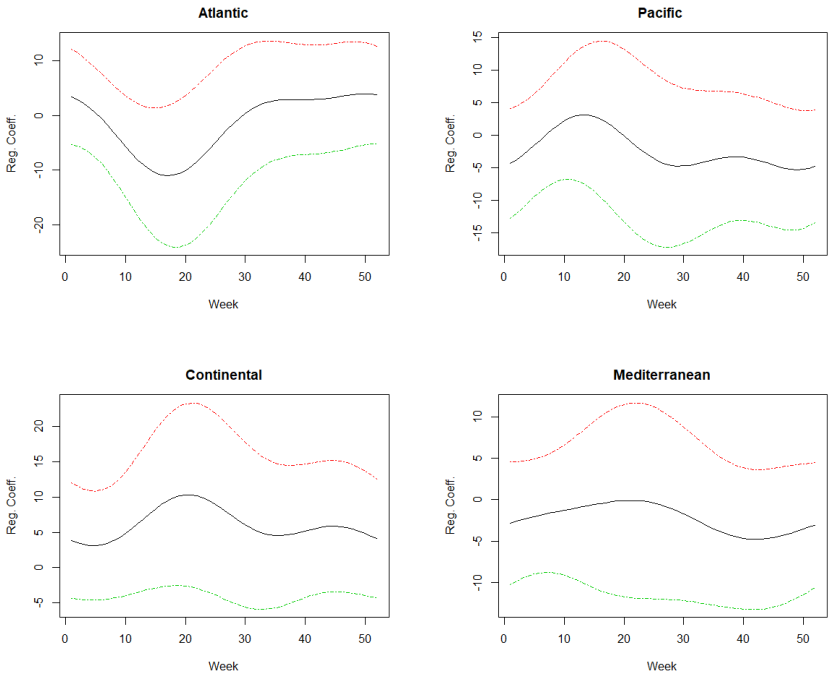


Figure 44. The region effects α_k for the humidity functions in the functional analysis of variance model. The effects $\alpha_k(t)$ are required to sum to 0 for all t. The dashed lines indicates 95% point-wise confidence intervals for the true effects.

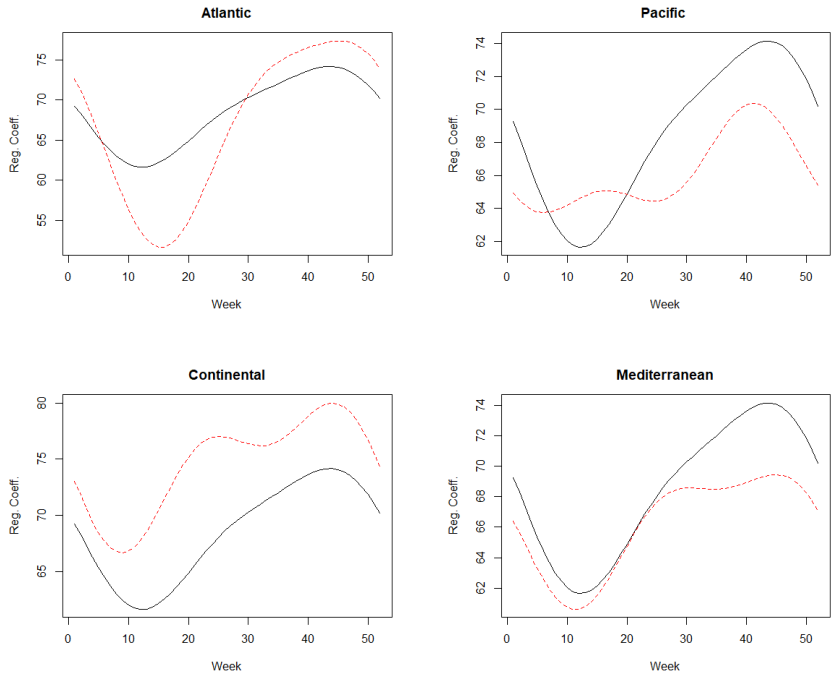


Figure 45. The dashed red curves are estimated climate zone humidity profile $\mu + \alpha_k$ for the humidity functions in the functional analysis of variance model. The solid black curve is the mean function μ of all cities.

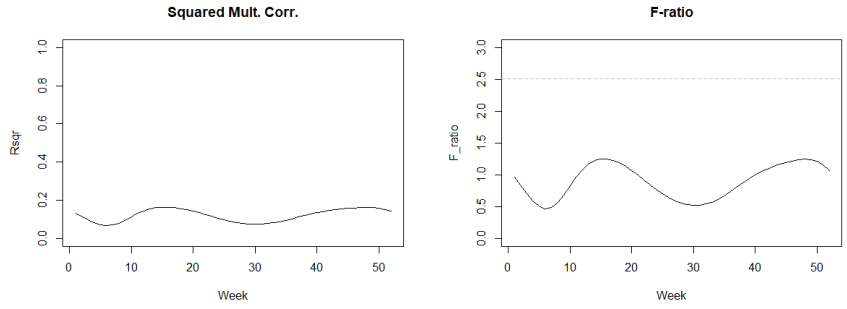


Figure 46. (i) Squared multiple correlation RSQ. (ii) The corresponding F-ratio function FRATIO. the dotted gray line indicates 5% significant level for the F-distribution with 5 and 32 degrees of freedom.

For variable pressure, $Pres_{ik}$ is the pressue of i-th city in the k-th climate group, we have the model of following form:

$$Pres_{ik}(t) = \mu(t) + \alpha_{ik}(t) + \varepsilon_{ik}(t).$$

Here α_k represents the specific effects on pressure in climate group k, with a constraint $\sum_k \alpha_k(t) = 0$ for all $t \in T$.

Figure 47 displays the estimated regions effects, along with 95% pointwise confidence intervals estimated. Figure 48 shows the composite effects $\mu + \alpha_k$. We see that the Atlantic coastal cities have a pressure around 10 hPa lower than the mean pressure of all cities. But Pacific coastal cities are hold contrary situation. The Pacific coastal cities have a higher pressure than mean level, but even more so in Spring. The continental cities hold the same situations as Pacific coastal cities that higher than the mean pressure, but more in Summer and Autumn. The Mediterranean cities pressure are very close to the mean average, but slightly lower in July and higher in Dec and Jan.

Results of RSQ and FRATIO functions are shown in Figure 49. The squared correlation is relatively high approaching to 1. F-ratio is everywhere higher than the 5% significant level for the F-distribution with 5 and 32 degrees of freedom, in which case is 2.51. The difference between climate zones are substantially stronger in the Winter than that in other seasons.

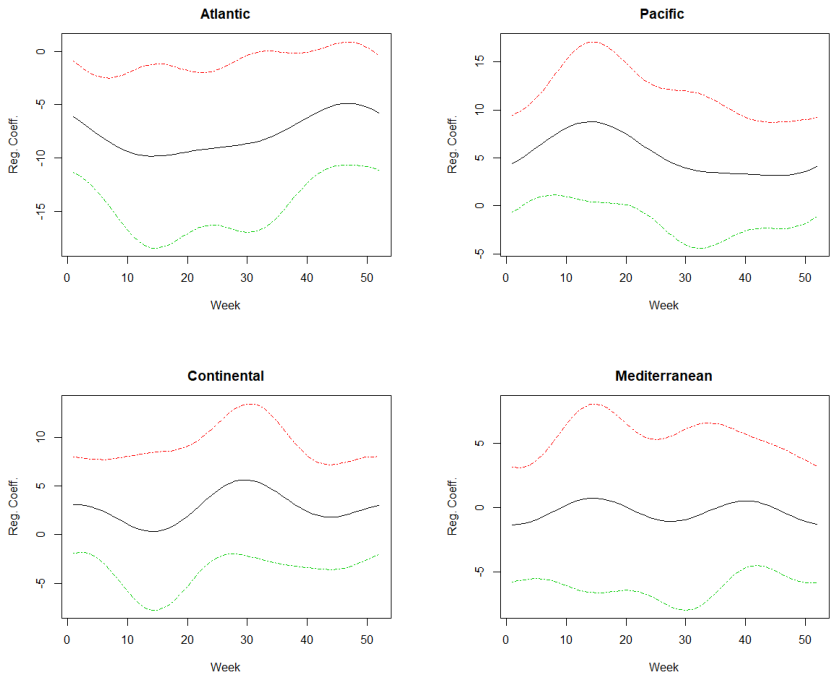


Figure 47. The region effects α_k for the pressure functions in the functional analysis of variance model. The effects $\alpha_k(t)$ are required to sum to 0 for all t. The dashed lines indicates 95% point-wise confidence intervals for the true effects.

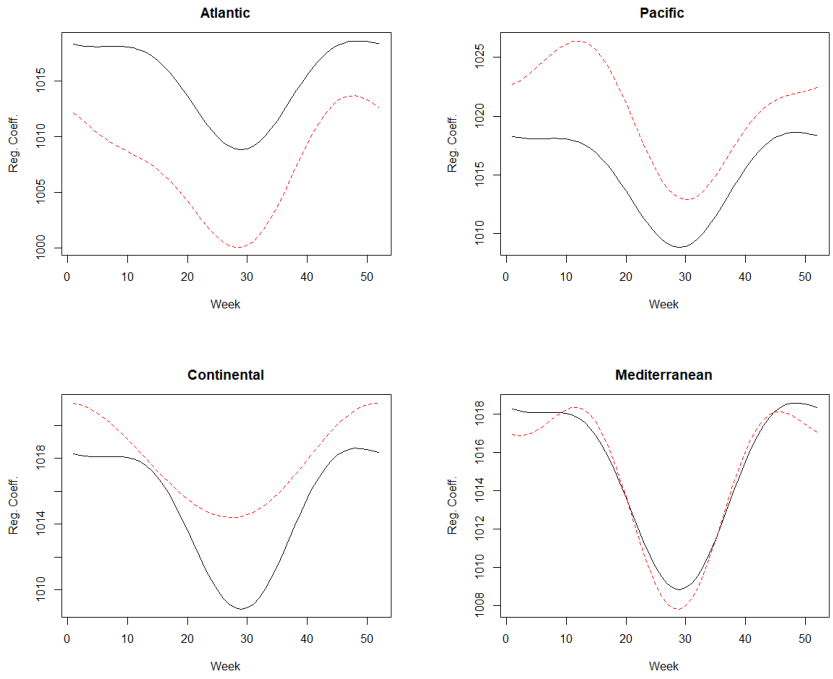


Figure 48. The dashed red curves are estimated climate zone pressure profile $\mu + \alpha_k$ for the pressure functions in the functional analysis of variance model. The solid black curve is the mean function μ of all cities.

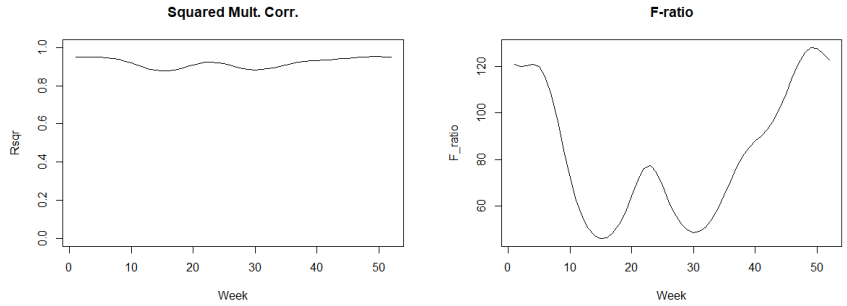


Figure 49. (i) Squared multiple correlation RSQ. (ii) The corresponding F-ratio function FRATIO.

For variable pressure, WS_{ik} is the pressue of i-th city in the k-th climate group, we have the model of following form:

$$WS_{ik}(t) = \mu(t) + \alpha_{ik}(t) + \varepsilon_{ik}(t).$$

Here α_k represents the specific effects on wind speed in climate group k, with a constraint $\sum_k \alpha_k(t) = 0$ for all $t \in T$.

Figure 50 displays the estimated regions effects on wind speed, along with 95% point-wise confidence intervals estimated. Figure 51 shows the composite effects $\mu + \alpha_k$. The Atlantic coastal cities wind speed is higher than the mean average of all cities, but much higher in Summer and Autumn. The pacific coastal cities are lower than the mean average wind speed around 1 unit. The continental cities hold the opposite situations as Pacific coastal cities. The wind speed of continental cities are higher than the mean average wind speed, but even more so in Winter and Spring. The Mediterranean cities wind speed are slightly fluctuate up and down upon the mean average.

Results of RSQ and FRATIO functions are shown in Figure 52. F-ratio is partially higher than the 5% significant level. From March to Jun, the F-ratio is lower than the 5% significant level. The difference between climate zones are substantially stronger in the Spring and Winter.

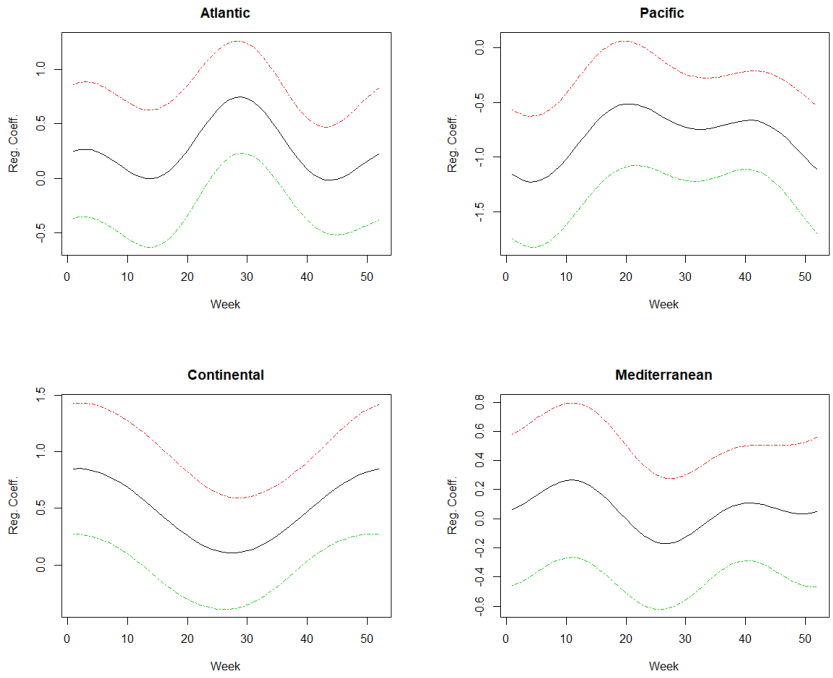


Figure 50. The region effects α_k for the wind speed functions in the functional analysis of variance model. The effects $\alpha_k(t)$ are required to sum to 0 for all t. The dashed lines indicates 95% point-wise confidence intervals for the true effects.

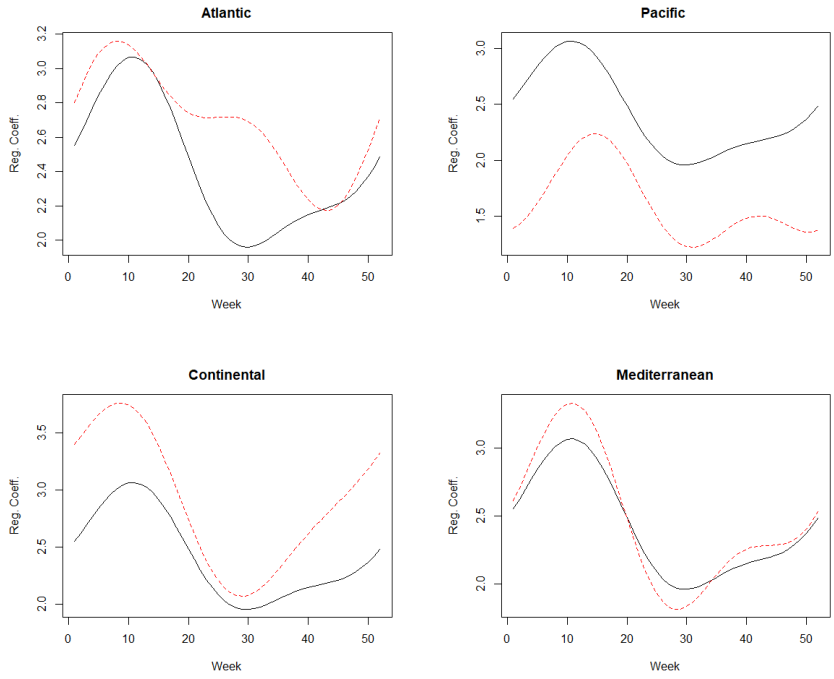


Figure 51. The dashed red curves are estimated climate zone wind speed profile $\mu + \alpha_k$ for the wind speed functions in the functional analysis of variance model. The solid black curve is the mean function μ of all cities.

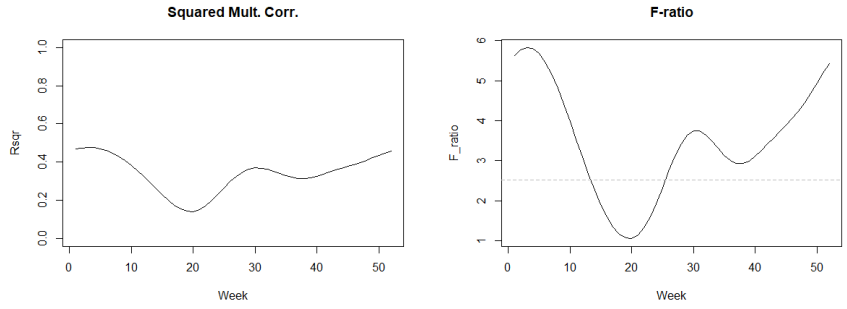


Figure 52. (i) Squared multiple correlation RSQ. (ii) The corresponding F-ratio function FRATIO. the dotted gray line indicates 5% significant level for the F-distribution with 5 and 32 degrees of freedom.

For variable pressure, WD_{ik} is the pressue of i-th city in the k-th climate group, we have the model of following form:

$$WD_{ik}(t) = \mu(t) + \alpha_{ik}(t) + \varepsilon_{ik}(t).$$

Here α_k represents the specific effects on wind direction in climate group k, with a constraint $\sum_k \alpha_k(t) = 0$ for all $t \in T$.

Figure 53 displays the estimated regions effects on wind direction, along with 95% pointwise confidence intervals estimated. Figure 54 shows the composite effects $\mu + \alpha_k$. The results of RSQ and FRATIO functions are shown in Figure 55. F-ratio is partially higher than the 5% significant level and the difference between climate regions are substantially stronger in the Summer and Winter.

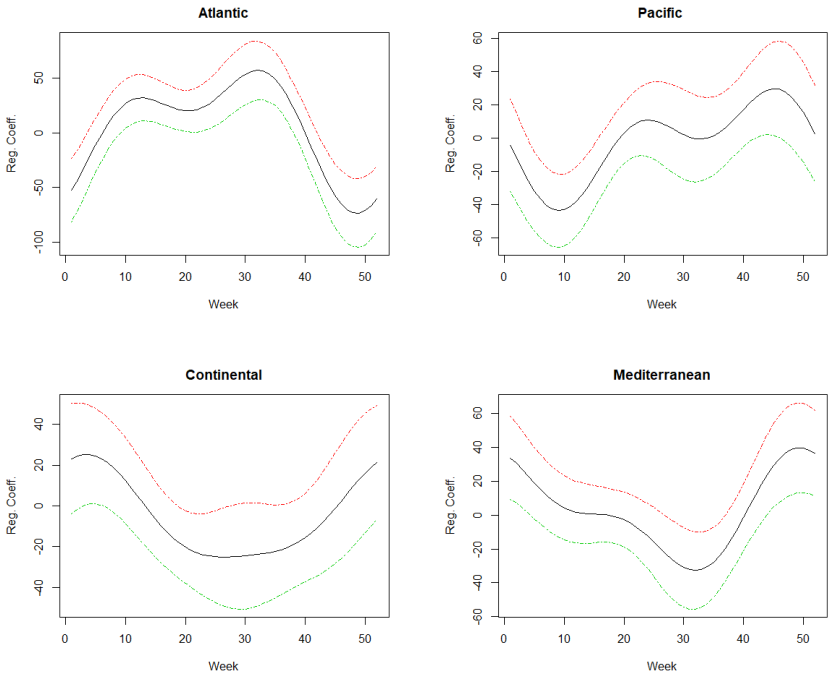


Figure 53. The region effects α_k for the wind direction functions in the functional analysis of variance model. The effects $\alpha_k(t)$ are required to sum to 0 for all t. The dashed lines indicates 95% point-wise confidence intervals for the true effects.

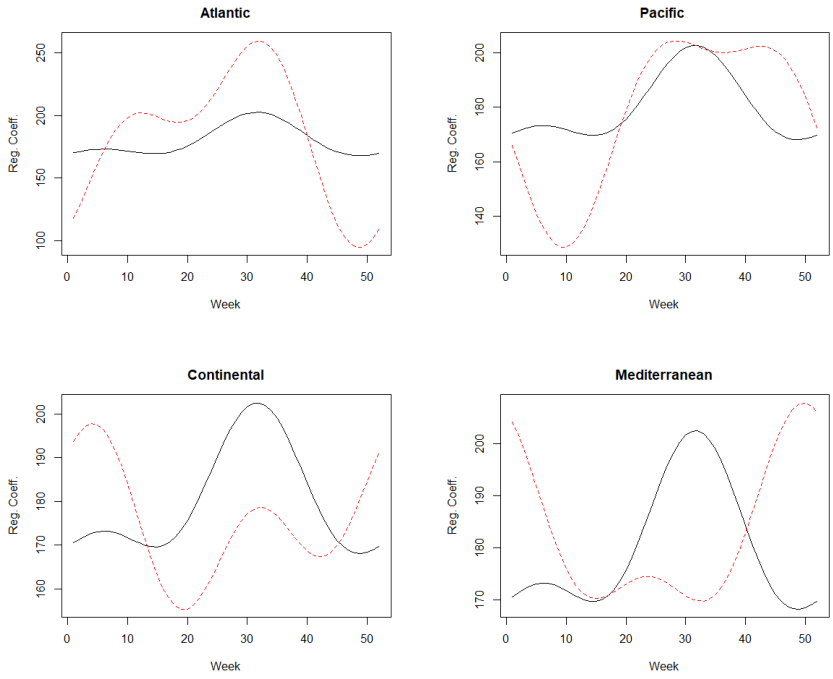


Figure 54. The dashed red curves are estimated climate zone wind direction profile $\mu + \alpha_k$ for the wind direction functions in the functional analysis of variance model. The solid black curve is the mean function μ of all cities.

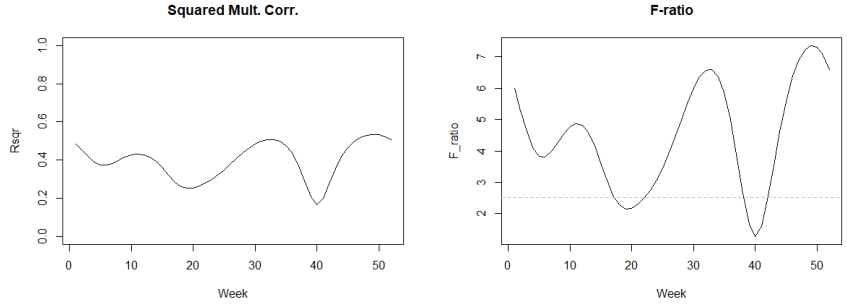


Figure 55. (i) Squared multiple correlation RSQ. (ii) The corresponding F-ratio function FRATIO. the dotted gray line indicates 5% significant level for the F-distribution with 5 and 32 degrees of freedom.

4.5.3 Fully Functional Regression Model

The aim of this subsection is to implement a fully functional linear model in which both response y and covariate z are functions. The aim of this section is fourfold. We investigated to what extent we can predict the complete weekly temperature profile of a city from information in its complete weekly other climate indicator profile. Then compare the results to see which climate indicator profile is better for predicting the temperature profile. We proposed the data by fitting a Fourier basis with 5 terms, applying a roughness penalty smoother, smoothing parameter λ was set by GCV.

The first pair of climate indicator is temperature and humidity, that is using the complete weekly humidity profile predict the temperature profile. So we have following functional linear model:

$$Temp_i(t) = \alpha(t) + \int_0^{52} Humi_i(t)\beta(s,t)ds + \varepsilon_i(t)$$

.

Considered the expression of $\beta(s,t)$ as double expansion with basis functions ϕ and ψ. We used the same 5 basis functions to expand the temperature and humidity functions for both basis system ϕ and ψ. Figure 56 displays the regression functions for the intercept and humidity effects. The surface 56(ii) shows the influence of humidity at time s on temperature at time t. The resulting prediction of the annual pattern of temperature at four randomly selected cities is demonstrated in Figure 57. We can find out some situations of the effect of humidity on temperature. Humidity from February to March is negatively associated with temperature throughout the year. Humidity from April to June is positively associated with temperature throughout the year. Then humidity from July to September is negatively associated with temperature throughout the entire year. Finally, humidity from Oct to Jan associated positively with temperature throughout the year, particularly with Summer temperature.

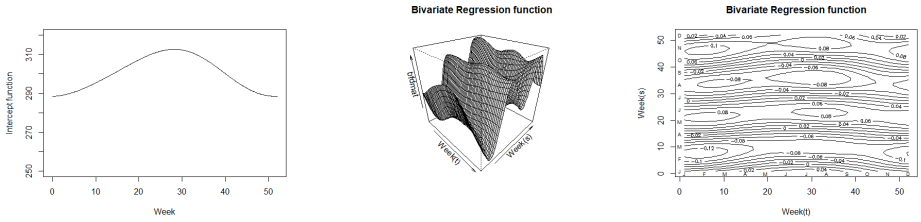


Figure 56. (i) Intercept coefficient. (ii) Perspective plot of estimated β function for the prediction of temperature from humidity, estimated direct from data. The value $\beta(s,t)$ shows the influence of humidity at time s on temperature at time t. (iii) Contour plot of corresponding estimated β function.

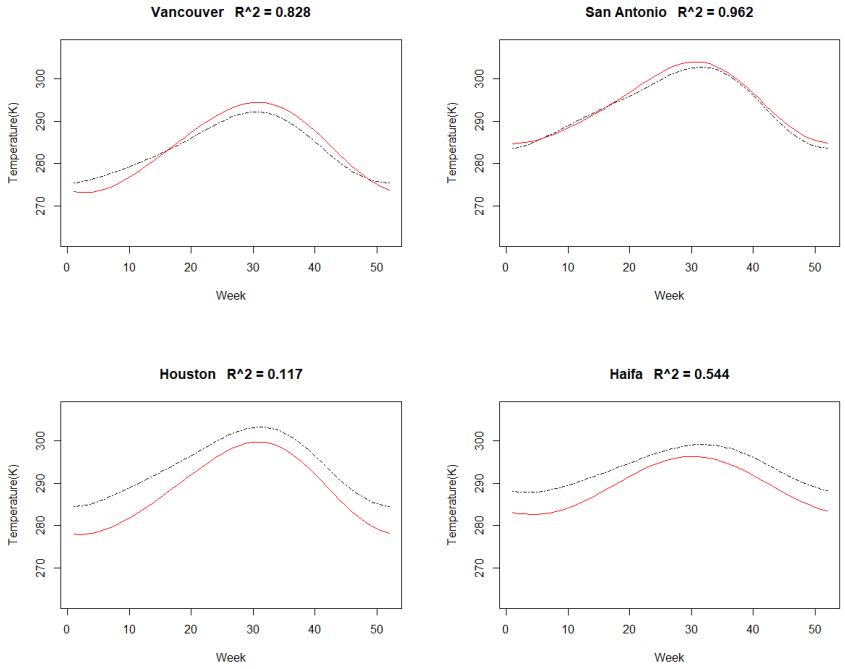


Figure 57. Original data(solid red line) and predictions(dashed black line) of the temperature relative to annual mean for each of 4 random selected cities.

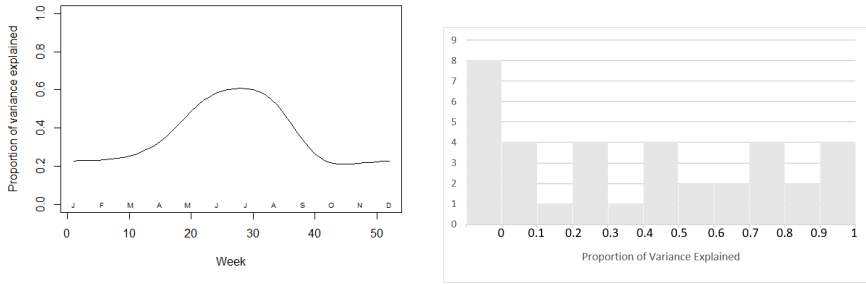


Figure 58. (i) Proportion of variance of temperature explained by a linear model based on weekly humidity records. (ii) Histogram of individual proportions of variance R_i^2 in temperature explained by a linear model based on weekly humidity records. The left-hand cell of the histogram includes all cases with negative R_i^2 values.

Figure 58 (i) plots the R^2 function for the fit to the temperature data. The fit is better in the Summer than other seasons. Figure 57 are overall R^2 measure for four randomly

selected cities. The value of R_i^2 are 0.828, 0.962, 0.117 and 0.544 respectively, clarifying that Vancouver and San Antonio are cities whose temperature fit closely to those predicted by the model based on their observed humidity profiles; for Haiva the fit is still not so bad in that the humidity pattern accounts for over 54% of the variation of the temperature from the overall population mean. Figure 58 (ii) gives a histogram of all involved 36 cities' R_i^2 values. We see that some cities, the R_i^2 value indicates excellent prediction, but for some cities the temperature pattern is not well predicted at all. There are 8 cities having negative R_i^2 values showing that for these cities the mean function actually provides a better fit to the true value than does the predictor.

Figure 59 plots the F-ratio function for the fit to the temperature profile. The upper 5% and 1% of the $F_{4:31}$ distribution are given. Within this model, F-ratio indicates that the effect of weekly humidity on temperature is highly significant during April to August.

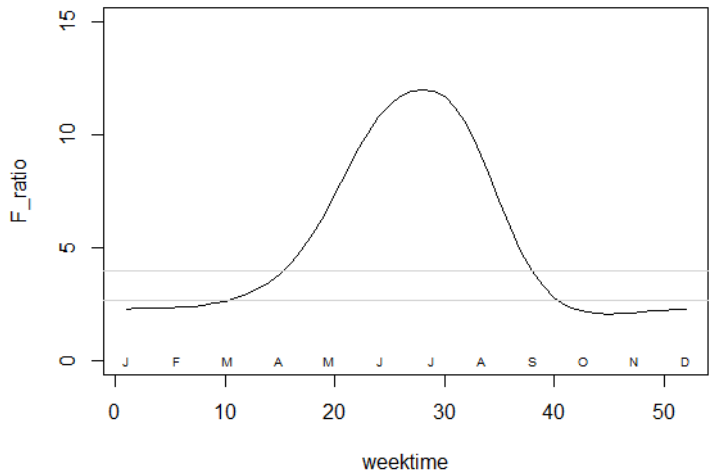


Figure 59. A plot of F-ratio function for the prediction of temperature from weekly humidity data. The prediction is carried out using an estimated β function with both bases are 5 terms. The horizontal gray lines show the upper 5% and 1% points of $F_{4:31}$ distribution.

The second scenario is predicting the complete weekly temperature profile from complete weekly pressure profile, so we have following functional linear model:

$$Temp_i(t) = \alpha(t) + \int_0^{52} Pres_i(t)\beta(s,t)ds + \varepsilon_i(t)$$

.

Figure 60 provides the intercept function and pressure regression function. The surface 60(ii) shows the effects of pressure at time s on temperature at time t. The resulting prediction of the annual pattern of temperature at four randomly selected cities is

demonstrated in Figure 61. Pressure from January to March is positively associated with temperature throughout the year. Pressure from April to June is negatively associated with temperature throughout the year. Then pressure from July to October is negatively associated with temperature in Spring and Winter. Finally, pressure from November to December associated positively with temperature throughout the year, particularly with winter temperature.

The R^2 function of the fitted temperature curve displayed in Figure 62 (i). The fit is poor. Figure 61 are overall R^2 measure for four randomly selected cities. The value of R_i^2 are 0.7, 0.364, 0.335 and 0.678 respectively, illustrating that Seattle and Vancouver whose temperature annual pattern explain over 60% of the variation of the temperature from the overall population mean. Figure 62(ii) provides a histogram of all 36 cities' R_i^2 values. We see that for most cities, the R_i^2 value are smaller than 0.5. There are 12 cities having negative R_i^2 values showing that for these cities the population mean function actually provides a better fit to the true value than does the predictor.

Figure 63 is the F-ratio function for the fit to the temperature data. The upper 5% and 1% of the $F_{4:31}$ distribution are given. Which shows that the effect of weekly pressure on temperature is not significant.

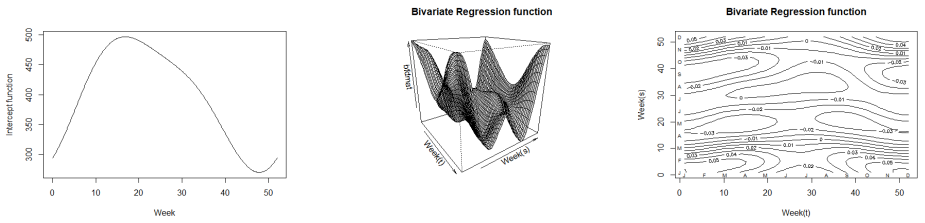


Figure 60. (i) Intercept coefficient. (ii) Perspective plot of estimated β function for the prediction of temperature from humidity, estimated direct from data. The value $\beta(s,t)$ shows the influence of humidity at time s on temperature at time t. (iii) Contour plot of estimated β function.

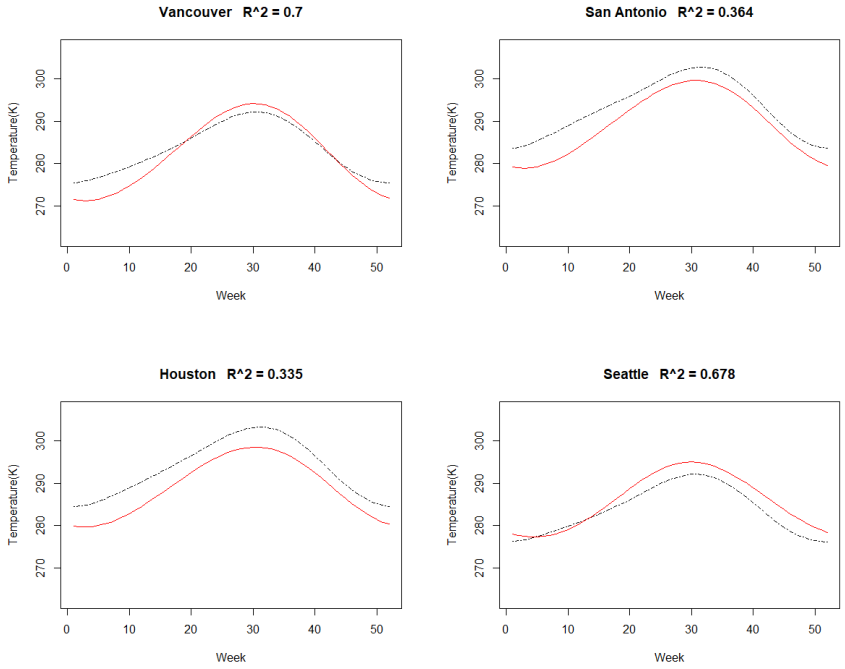


Figure 61. Original data(solid red line) and predictions(dashed black line) of the temperature relative to annual mean for each of 4 random selected cities.

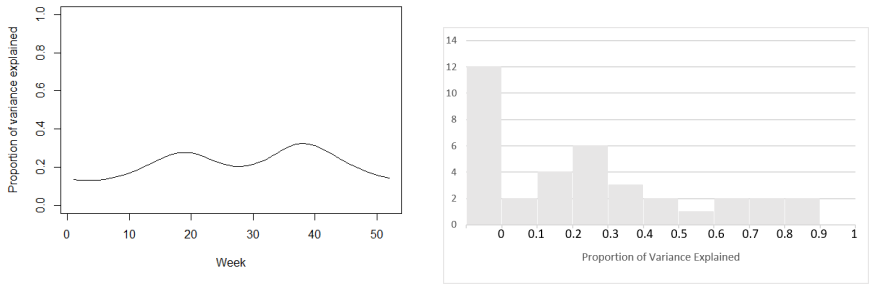


Figure 62. (i) Proportion of variance of temperature explained by a linear model based on weekly humidity records. (ii) Histogram of individual proportions of variance R_i^2 in temperature explained by a linear model based on weekly humidity records. The left-hand cell of the histogram includes all cases with negative R_i^2 values.

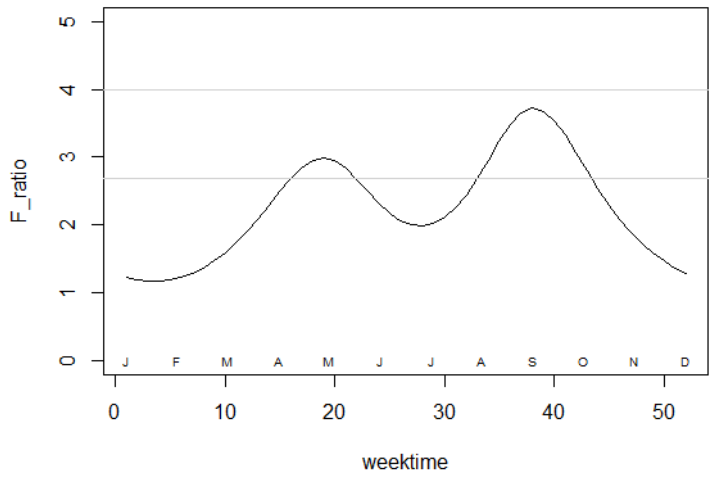


Figure 63. A plot of F-ratio function for the prediction of temperature from weekly pressure data. The prediction is carried out using an estimated β function with both bases are 5 terms. The horizontal gray lines show the upper 5% and 1% points of $F_{4:31}$ distribution.

The third pair of climate indicator is temperature and wind speed, so we have following functional linear model:

$$Temp_i(t) = \alpha(t) + \int_0^{52} WS_i(t)\beta(s,t)ds + \varepsilon_i(t)$$

.

The regression functions of intercept the pressure effects are shown in Figure 64. The surface 64(ii) shows the influence of pressure at time s on temperature at time t. The resulting prediction of the annual pattern of temperature at four random selected cities is demonstrated in Figure 65. Wind speed from January to February is positively associated with temperature throughout the year, particularly with winter temperature. From March to May, the wind speed is negatively associated with temperature throughout the year. Then wind speed from May to Mid-Autumn is positively associated with temperature throughout the year. Finally, wind speed from Mid-Autumn to December, the wind speed is negatively associated with temperature.

Figure 66 (i) gives the R^2 function for the fit to the temperature. Figure 65 are overall R^2 measure for four randomly selected cities. The value of R_i^2 are 0.959, 0.987, 0.986 and 0.784 respectively, illustrating that all four selected cities are places whose temperature fit closely to those predicted by the model based on their observed wind speed profiles. Figure 66(ii) is the histogram gives all involved 36 cities' R_i^2 values. We see that for most of the cities, the R_i^2 value indicates excellent predictions, but for a small proportion the temperature pattern is not at all well predicted. There are 7 cities having negative R_i^2

values, which means for these cities, the population mean actually provides a better fit to the true value than does the predictor.

Figure 67 plots the F-ratio function for the fit to the temperature data. The upper 5% and 1% of the $F_{4:31}$ distribution are given. We see that the effect of weekly wind speed on temperature is partially highly significant, particularly in Spring.

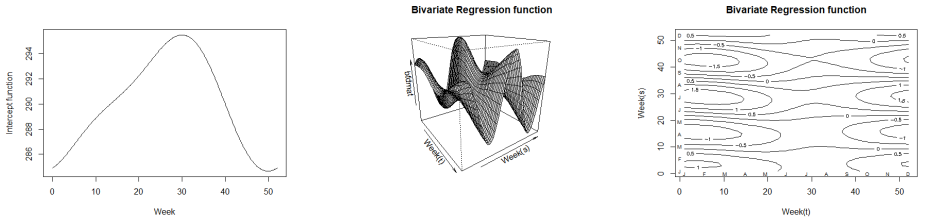


Figure 64. (i) Intercept coefficient. (ii) Perspective plot of estimated β function for the prediction of temperature from wind speed, estimated direct from data. The value $\beta(s,t)$ shows the influence of wind speed at time s on temperature at time t. (iii) Contour plot of estimated β function.

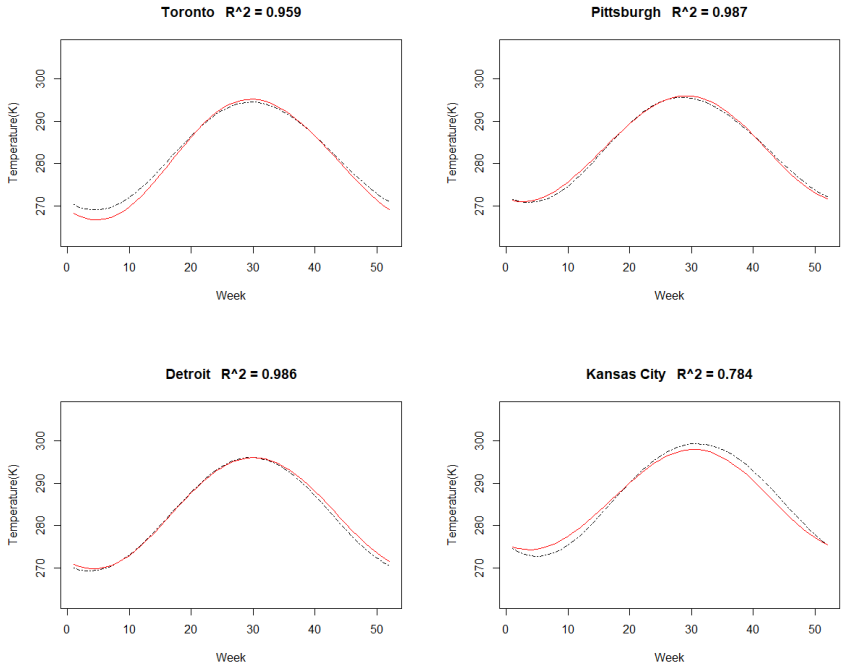


Figure 65. Original data(solid red line) and predictions(dashed black line) of the temperature relative to annual mean for each of 4 random selected cities.

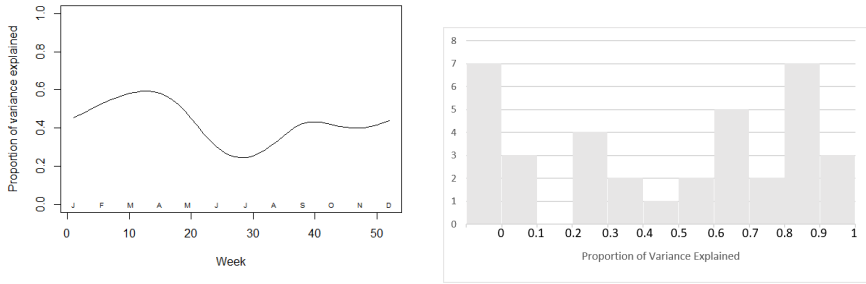


Figure 66. (i) Proportion of variance of temperature explained by a linear model based on weekly humidity records. (ii) Histogram of individual proportions of variance R_i^2 in temperature explained by a linear model based on weekly humidity records. The left-hand cell of the histogram includes all cases with negative R_i^2 values.

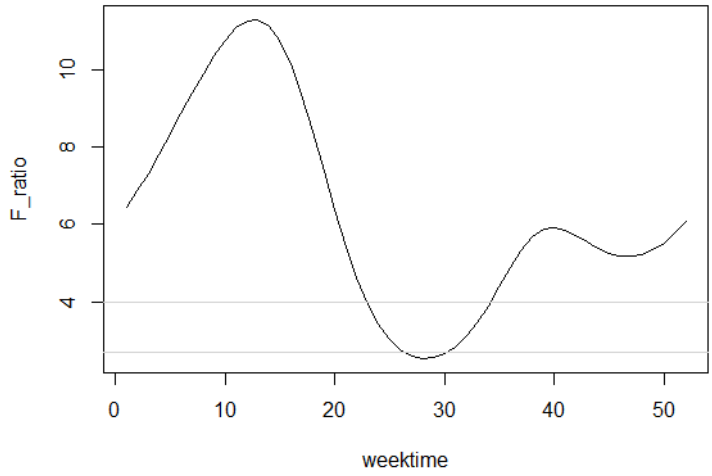


Figure 67. A plot of F-ratio function for the prediction of temperature from weekly pressure data. The prediction is carried out using an estimated β function with both bases are 5 terms. The horizontal gray lines show the upper 5% and 1% points of $F_{4:31}$ distribution.

The last pair of climate indicator is temperature and wind direction, so we have following functional linear model:

$$Temp_i(t) = \alpha(t) + \int_0^{52} WD_i(t)\beta(s,t)ds + \varepsilon_i(t)$$

.

Figure 68 gives the regression functions for the intercept and wind direction effects. The surface 68 (ii) reflects the effect of wind direction at time s on temperature at time t. Figure 69 gives the prediction of annual temperature at four randomly selected cities. Wind direction from March to June is positively associated with temperature throughout the year. Then Wind direction from July to November is negatively associated with temperature throughout the year.

Figure 70 (i) is the R^2 function for the fit to the temperature data. The fit is quite good. Figure 69 are overall R^2 measure for four randomly selected cities. The value of R_i^2 are 0.922, 0.961, 0.889 and 0.736 respectively, illustrating that these are the cities whose temperature fit closely to those predicted temperature. Figure 70(ii) is the histogram of all 36 involved cities' R_i^2 values. For most of the cities, the R_i^2 value are higher than 0.5 indicating excellent predictions. There are 6 cities having negative R_i^2 values. For these cities, the population mean actually provides a better fit to the true value than does the predictor.

Figure 71 is the F-ratio function for the fit to the temperature data. The upper 5% and 1% of the $F_{4:31}$ distribution are given. Within this model, the effect of weekly wind direction on temperature is highly significant overall.

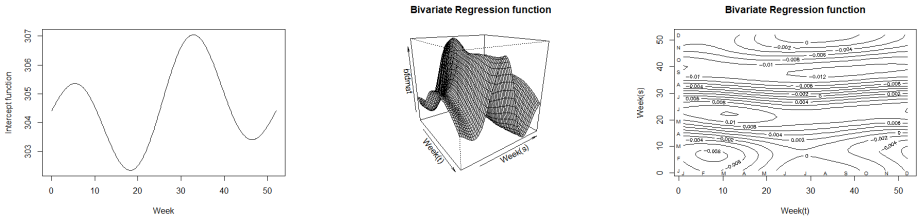


Figure 68. (i) Intercept coefficient. (ii) Perspective plot of estimated β function for the prediction of temperature from wind direction, estimated direct from data. The value $\beta(s,t)$ shows the influence of humidity at time s on temperature at time t. (iii) Contour plot of estimated β function.

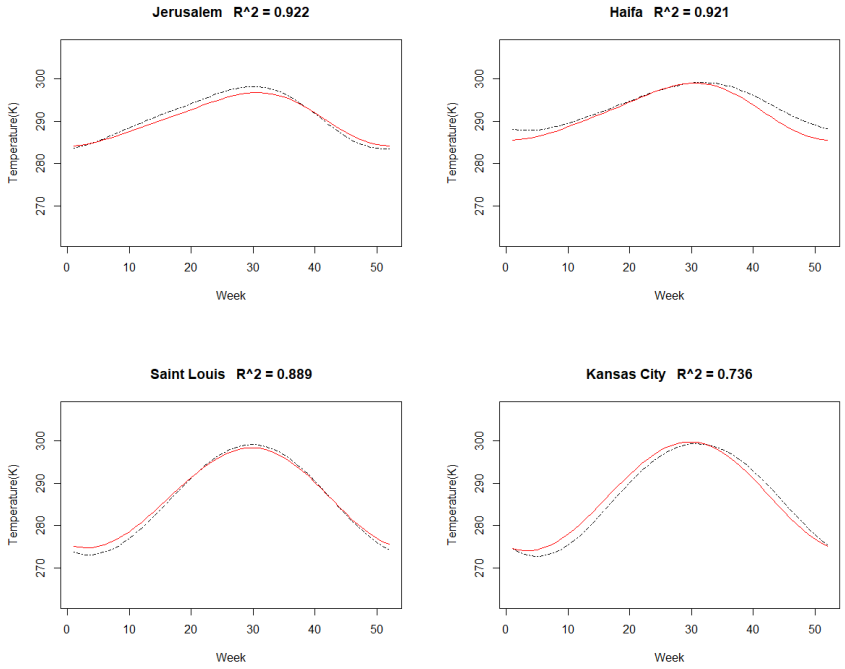


Figure 69. Original data(solid red line) and predictions(dashed black line) of the temperature relative to annual mean for each of 4 random selected cities.

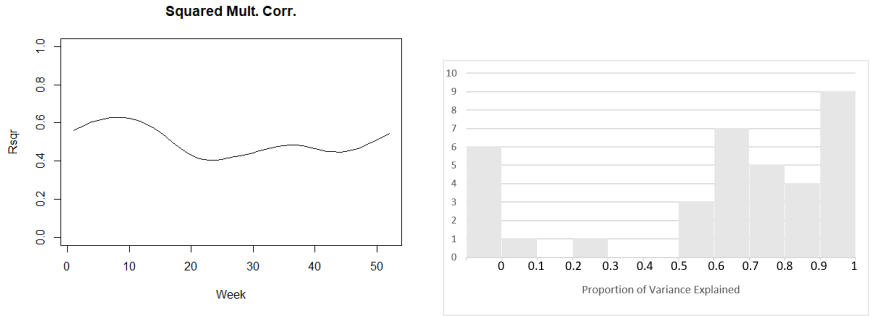


Figure 70. (i) Proportion of variance of temperature explained by a linear model based on weekly humidity records. (ii) Histogram of individual proportions of variance R_i^2 in temperature explained by a linear model based on weekly humidity records. The left-hand cell of the histogram includes all cases with negative R_i^2 values.

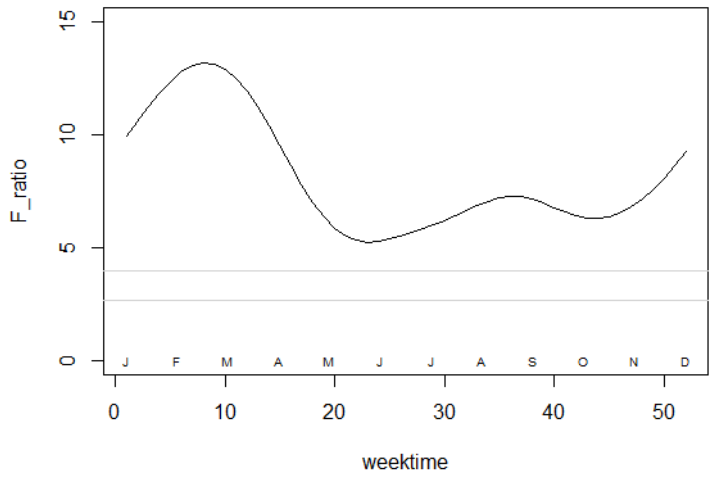


Figure 71. A plot of F-ratio function for the prediction of temperature from weekly pressure data. The prediction is carried out using an estimated β function with both bases are 5 terms. The horizontal gray lines show the upper 5% and 1% points of $F_{4:31}$ distribution.

5 Conclusion

This Chapter will provide an overview and summaries subject to this thesis. A briefly summary of the illustration implemented in Chapter 4 is also provided in this Chapter. Also, some recommendations will be made for further research regarding Functional Data Analysis or Functional Linear Regression Modelling.

5.1 Concluding Remarks

- Chapter 2 was to define the *Functional Data Analysis* and introduce some important basis functions and smoothing techniques that were used throughout this thesis. An In-depth explanation of the basis expansions is provided. In this thesis, the emphasis was on the relevant methods, namely B-spline Bases and Fourier Bases. The roughness penalty method is introduced to produce a better estimation. The GCV criteria is used to locate a best parameter for smoothing. Some definition of functional descriptive statistics were also introduced in Chapter 2 and relevant interpretation were provided in Chapter 4 and Appendix.

- Another main part of Chapter 2 was providing the mathematical foundations of *Functional Principal Component Analysis*. The relevant background to the meaning of the *Karhunen-Loeve* expansion was also given. And two graphical methods

based on ordering the first two robust principal component scores are introduced for detecting outliers. Their relevant interpretation provided in Chapter 4 and Appendix.

- In Chapter 3, the *Functional variance analysis* and *Fully Functional Linear Regression* model were introduced. The goodness of fit evaluation is consider the squared correlation function and F-ratio function.

- Chapter 4 applied all the FDA methods introduced in this thesis on a real climate data set. The B-spline and Fourier basis were used to smooth weekly temperature, humidity, pressure, wind speed and wind direction for all 36 cities. GCV was used to compute optimal roughness penalty parameters.

  The fPCA and *Functional Linear Regression* model were implemented based on smoothing curves using Fourier basis with 5 terms, and *Roughness Penalty* parameter set by *Generalized Cross-Validation* (GCV). Once the functional data of temperature were obtained, the functional linear regression with one univariate predictor was implemented for prediction of the temperature functional profile (See Figures 57, 61, 65, 69). The R^2 and F-ratio function were computed for the goodness-of-fit. Based on results, using wind direction profile to predict the temperature file has better results than using other climate indicator profiles to predict the temperature profile.

5.2 Recommendations

The **R**-package **fda** by Ramsay et al. (2009) [46] is the most popular and useful package for functional data analysis. Unfortunately, the package **fda** has a bit restricted. For example, when it comes to *Functional Linear Regression* model, R function *fRegress* carries out a functional regression analysis, where the either the response or one or more predictors are functional. However, only univariate independent variables are currently allowed. More packages or functions must be released in that regard.

Regarding *Functional Linear Regression* modelling, a vast area of research is still worth to explore. This thesis only provides the first step in this direction.

References

[1] A. M. Aguilera and M. C. Aguilera-Morillo. Comparative study of different b-spline approaches for functional data. *Mathematical and computer modelling*, 58(7-8):1568–1579, 2013.

[2] Ana M. Aguilera, Manuel Escabias, and Mariano J. Valderrama. Discussion of different logistic models with functional data. application to systemic lupus erythematosus. *Computational Statistics & Data Analysis*, 53(1):151 – 163, 2008.

[3] John A. D. Aston, Jeng-Min Chiou, and Jonathan P. Evans. Linguistic pitch analysis using functional principal component mixed effect models. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 59(2):297–317, 2010.

[4] Fateh Chebana, Sophie Dabo-Niang, and Taha BMJ Ouarda. Exploratory functional flood frequency analysis and outlier detection. *Water Resources Research*, 48(4), 2012.

[5] Yizhi Chen. Function data analysis method and application(in chinese). *Hangzhou: Zhejiang University of Industry and Commerce*, 2011.

[6] Mariarosaria Coppola, Maria Russolillo, and Rosaria Simone. An indexation mechanism for retirement age: Analysis of the gender gap. *Risks*, 7(1):1–13, 2019.

[7] Antonio Cuevas, Manuel Febrero, and Ricardo Fraiman. An anova test for functional data. *Computational Statistics & Data Analysis*, 47(1):111 – 122, 2004.

[8] Stelian Curceac, Camille Ternynck, Taha B.M.J. Ouarda, Fateh Chebana, and Sophie Dabo Niang. Short-term air temperature forecasting using nonparametric functional data analysis and sarma models. *Environmental Modelling & Software*, 111:394 – 408, 2019.

[9] Carl De Boor. On calculating with b-splines. *Journal of Approximation theory*, 6(1):50–62, 1972.

[10] Carl De Boor, Carl De Boor, Etats-Unis Mathématicien, Carl De Boor, and Carl De Boor. *A practical guide to splines*, volume 27. springer-verlag New York, 1978.

[11] Rene Franck Essomba. An investigation into functional linear regression modeling. Master's thesis, University of Cape Town, 2015.

[12] Frédéric Ferraty and Philippe Vieu. Additive prediction and boosting for functional data. *Computational Statistics & Data Analysis*, 53(4):1400 – 1413, 2009.

[13] Frédéric Ferraty and Philippe Vieu. *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media, 2006.

[14] Roger G Ghanem and Pol D Spanos. *Stochastic finite elements: a spectral approach*. Courier Corporation, 2003.

[15] Tomasz Górecki and Łukasz Smaga. fdanova: an r software package for analysis of variance for univariate and multivariate functional data. *Computational Statistics*, 34(2):571–597, 2019.

[16] Peter J Green and Bernard W Silverman. *Nonparametric regression and generalized linear models: a roughness penalty approach*. Crc Press, 1993.

[17] Siegfried Hörmann, Łukasz Kidziński, and Marc Hallin. Dynamic functional principal components. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 319–348, 2015.

[18] Lajos Horváth and Piotr Kokoszka. *Inference for Functional Data with Applications*. Springer Series in Statistics. Springer-Verlag New York, 1 edition, 2012.

[19] Lixia Hu, Tao Huang, and Jinhong You. Unified statistical inference for a novel nonlinear dynamic functional/longitudinal data model. *arXiv preprint arXiv:2007.01784*, 2020.

[20] Rob J. Hyndman. Computing and graphing highest density regions. *The American Statistician*, 50(2):120–126, 1996.

[21] Rob J. Hyndman and Md Shahid Ullah. Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis*, 51(10):4942–4956, 2007.

[22] Rob J. Hyndman and Han Lin Shang. Rainbow plots, bagplots, and boxplots for functional data. *Journal of Computational and Graphical Statistics*, 19(1):29–45, 2010.

[23] Takayoshi Ikeda, Michael Dowd, and Jennifer L. Martin. Application of functional data analysis to investigate seasonal progression with interannual variability in plankton abundance in the bay of fundy, canada. *Estuarine, Coastal and Shelf Science*, 78(2):445–455, 2008.

[24] Meredith C. King, Ana-Maria Staicu, Jerry M. Davis, Brian J. Reich, and Brian Eder. A functional data analysis of spatiotemporal trends and variation in fine particulate matter. *Atmospheric Environment*, 184:233–243, 2018.

[25] DD Kosambi. *Statistics in function space*, pages 115–123. Springer, 2016.

[26] Nicole Krämer, Anne-Laure Boulesteix, and Gerhard Tutz. Penalized partial least squares with applications to b-spline transformations and functional data. *Chemometrics and Intelligent Laboratory Systems*, 94(1):60–69, 2008.

[27] Algirdas Laukaitis and Alfredas Račkauskas. Functional data analysis for clients segmentation tasks. *European Journal of Operational Research*, 163(1):210–216, 2005.

[28] Sara López-Pintado and Juan Romo. A half-region depth for functional data. *Computational Statistics & Data Analysis*, 55(4):1679 – 1695, 2011.

[29] J. Lucero. Computation of the harmonics-to-noise ratio of a voice signal using a functional data analysis algorithm. *Sound and Vibration*, 222:512–520, 1999.

[30] Samuel Maistre and Valentin Patilea. Testing for the significance of functional covariates. *Journal of Multivariate Analysis*, 179:104648, 2020.

[31] Javier Martínez Torres, Jorge Pastor Pérez, Joaquín Sancho Val, Aonghus McNabola, Miguel Martínez Comesaña, and John Gallagher. A functional data analysis approach for the detection of air pollution episodes and outliers: A case study in dublin, ireland. *Mathematics*, 8(2), 2020.

[32] Hidetoshi Matsui, Shuichi Kawano, and Sadanori Konishi. Regularized functional regression modeling for functional response and predictors. *Journal of Math-for-industry*, 1(3):17–25, 2009.

[33] J. S. Morris, C. Arroyo, B. A. Coull, L. M. Ryan, R. Herrick, and S. L. Gortmaker. Using wavelet-based functional mixed models to characterize population heterogeneity in accelerometer profiles: A case study. *J Am Stat Assoc*, 101(476):1352–1364, 2006.

[34] Stanislav Nagy. Statistical depth for functional data, 2016.

[35] H. Oliver Gao Niemeier and Debbie A. Using functional data analysis of diurnal ozone and nox cycles to inform transportation emissions control. *Transportation Research Part D: Transport and Environment*, 13:221–238, 2008.

[36] Juhyun Park, Theo Gasser, and Valentin Rousson. Structural components in functional data. *Computational Statistics & Data Analysis*, 53(9):3452–3465, 2009.

[37] J. S. Phillips, T. A. Patterson, B. Leroy, G. M. Pilling, and S. J. Nicol. Objective classification of latent behavioral states in bio-logging data using multivariate-normal hidden markov models. *Ecol Appl*, 25(5):1244–58, 2015.

[38] A. Pourshoghi, I. Zakeri, and K. Pourrezaei. Application of functional data analysis in classification and clustering of functional near-infrared spectroscopy signal in response to noxious stimuli. *J Biomed Opt*, 21(10):101411, 2016.

[39] Wilmer Prentius. Exploring cumulative incomefunctions by functional data analysis, 2016.

[40] Hadjipantelis P.Z. and Müller HG. Functional data analysis for big data: A case study on california temperature trends. *Springer Handbooks of Computational Statistics*, 2019.

[41] J. O. Ramsay. When the data are functions. *Psychometrika*, 47(4):379–396, 1982.

[42] J. O. Ramsay, Giles Hooker, and Spencer Graves. *Functional Data Analysis with R and MATLAB*. Springer Publishing Company, Incorporated, 1st edition, 2009.

[43] James Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer-Verlag New York, 1 edition, 1997.

[44] James O Ramsay and CJ Dalzell. Some tools for functional data analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(3):539–561, 1991.

[45] James O Ramsay and Bernard W Silverman. *Applied functional data analysis: methods and case studies*. Springer, 2007.

[46] JO Ramsay, Hadley Wickham, Maintainer JO Ramsay, and Suggests deSolve. Package 'fda'. 2020.

[47] Sarah J Ratcliffe, Gillian Z Heller, and Leo R Leader. Functional data analysis with application to periodically stimulated foetal heart rate data. ii: Functional logistic regression. *Statistics in medicine*, 21(8):1115–1127, 2002.

[48] Sarah J Ratcliffe, Leo R Leader, and Gillian Z Heller. Functional data analysis with application to periodically stimulated foetal heart rate data. i: Functional regression. *Statistics in Medicine*, 21(8):1103–1114, 2002.

[49] Jeffrey N. Rouder, Richard D. Morey, Paul L. Speckman, and Jordan M. Province. Default bayes factors for anova designs. *Journal of Mathematical Psychology*, 56(5):356 – 374, 2012.

[50] Peter J Rousseeuw and Ida Ruts. Algorithm as 307: Bivariate location depth. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 45(4):516–526, 1996.

[51] Peter J Rousseeuw, Ida Ruts, and John W Tukey. The bagplot: a bivariate boxplot. *The American Statistician*, 53(4):382–387, 1999.

[52] Stefania Salvatore. Application of functional data analysis (fda) to weekly wastewater data. 2017.

[53] David W Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

[54] N. Shaadan, S. M. Deni, and Abdul Aziz Jemain. Assessing and comparing pm10 pollutant behaviour using functional data approach. *Sains Malaysiana*, 41:1335–1344, 2012.

[55] Hanlin Shang et al. rainbow: an r package for visualizing functional time series. 2011.

[56] J.O. Ramsay Silverman and B.W. *Functional Data Analysis*. Springer-Verlag New York, 2005.

[57] Joon Jin Song, Weiguo Deng, Ho-Jin Lee, and Deukwoo Kwon. Optimal classification for time-course gene expression data using functional data analysis. *Computational biology and chemistry*, 32(6):426–432, 2008.

[58] Ana-Maria Staicu, Yingxing Li, Ciprian M. Crainiceanu, and David Ruppert. Likelihood ratio tests for dependent data with applications to longitudinal and functional data analysis. *Scandinavian Journal of Statistics*, 41(4):932–949, 2014.

[59] John W Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, volume 2, pages 523–531, 1975.

[60] Shahid Ullah and Caroline F Finch. Applications of functional data analysis: A systematic review. *BMC medical research methodology*, 13(1):43, 2013.

[61] Mariano J Valderrama. An overview to modelling functional data, 2007.

[62] Isaac Michael Wagner-Muns, Ivan G. Guardiola, V. A. Samaranayke, and Wasim Irshad Kayani. A functional data analysis approach to traffic volume forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 19(3):878–888, 2018.

[63] Grace Wahba. *Spline models for observational data*. SIAM, 1990.

[64] Deqing Wang. Research on statistical classification methods of functional data mining. *Doctoral dissertation. Xiamen: Xiamen University*, 2014.

[65] Jane-Ling Wang, Jeng-Min Chiou, and Hans-Georg Müller. Functional data analysis. *Annual Review of Statistics and Its Application*, 3:257–295, 2016.

[66] Zhiliang Wang, Yalin Sun, and Peng Li. Functional principal components analysis of shanghai stock exchange 50 index. *Discrete Dynamics in Nature and Society*, 2014, 2014.

[67] Farah Yasmeen, Rob J Hyndman, and Bircan Erbas. Forecasting age-related changes in breast cancer mortality among white and black us women: A functional data approach. *Cancer epidemiology*, 34(5):542–549, 2010.

[68] Jin-Ting Zhang. Statistical inferences for linear models with functional responses. *Statistica Sinica*, pages 1431–1451, 2011.

[69] Jin-Ting Zhang. *Analysis of variance for functional data*. CRC Press, 2013.

[70] Yuanyuan Zhang, Chienkai Wang, Fangfang Wu, Kun Huang, Lijian Yang, and Linhong Ji. Prediction of working memory ability based on eeg by functional data analysis. *Journal of Neuroscience Methods*, 333:108552, 2020.

[71] Hongxiao Zhu, Philip J Brown, and Jeffrey S Morris. Robust, adaptive functional regression in functional mixed model framework. *Journal of the American Statistical Association*, 106(495):1167–1179, 2011.

Appendices

Table 6. Summary statistics for air pressure of the second sample(weekly records)

City	Mean	SD	Median	City	Mean	SD	Median
Vancouver	1031.4165	22.239211	1090.932	Portland	1019.7541	6.952133	1038.5
San.Francisco	1018.3649	6.077586	1043	Seattle	1026.5128	6.580413	1037.963
Los.Angeles	1015.3962	5.650837	1039.5	San.Diego	1017.4041	5.843683	1038.5
Las.Vegas	1013.6635	8.344385	1044	Phoenix	1008.573	9.364447	1039.5
Albuquerque	1000.5354	25.841059	1042.5	Denver	1009.2733	23.704476	1045.232
San.Antonio	1015.6733	8.935095	1039.5	Dallas	1016.8727	6.508057	1039
Houston	1018.1443	5.768215	1038	Kansas.City	1006.7169	26.544626	1036
Minneapolis	1017.47	5.907473	1035.5	Saint.Louis	1017.3486	5.701971	1031.103
Chicago	1017.2595	5.133152	1032.667	Nashville	1018.4818	5.250517	1030.726
Indianapolis	1017.613	6.930368	1029.619	Atlanta	1019.8575	4.579584	1031.905
Detroit	1017.8522	4.846123	1031.714	Jacksonville	1019.5008	5.077366	1035.852
Charlotte	1019.4956	4.888808	1032.524	Miami	1019.4895	4.41419	1035.037
Pittsburgh	1018.4281	5.162766	1032.381	Toronto	1016.5045	5.101506	1030.238
Philadelphia	1019.1376	5.66031	1033.476	New.York	1013.8592	10.107746	1031
Montreal	1017.6404	5.751049	1032	Boston	1018.2256	5.01317	1030.889
Beersheba	988.0371	15.361048	1031	Tel.Aviv.District	1013.9259	5.354285	1030.5
Eilat	1006.402	9.542018	1028.5	Haifa	1016.8818	5.783307	1030
Nahariyya	1016.4298	5.662558	1030	Jerusalem	1001.9811	11.135762	1030

Table 7. Summary statistics for wind speed of the second sample(weekly records)

City	Mean	SD	Median	City	Mean	SD	Median
Vancouver	1.3506904	0.6269651	2.714286	Portland	1.4770531	0.6990675	4.388889
San.Francisco	2.1495164	1.1073573	5.005952	Seattle	1.6921174	0.5957153	3.35119
Los.Angeles	0.8147688	0.2021461	1.452381	San.Diego	1.5469328	0.488266	2.77381
Las.Vegas	2.1715083	0.9978824	4.678571	Phoenix	1.152927	0.3491021	2.10119
Albuquerque	2.4840271	0.8323809	5.005952	Denver	1.7727065	0.8005295	3.482143
San.Antonio	2.6038314	0.822513	4.089286	Dallas	2.9598864	1.0090241	5.434524
Houston	2.8148589	0.9952568	4.880952	Kansas.City	3.0205194	0.8264794	4.869048
Minneapolis	2.9835089	0.7809965	4.559524	Saint.Louis	2.4340127	0.8761417	5.297619
Chicago	3.0500351	1.1980436	6.916667	Nashville	2.2055192	0.9631167	4.392857
Indianapolis	2.9852395	1.0771913	5.821429	Atlanta	1.982489	0.6631466	3.410714
Detroit	2.9457641	0.783143	4.952381	Jacksonville	2.6095788	0.7406071	4.190476
Charlotte	1.8768882	0.6261893	4.113095	Miami	2.6942221	0.7905506	4.261905
Pittsburgh	2.4403559	1.0132395	4.297619	Toronto	3.1272002	1.5135311	6.638889
Philadelphia	2.4698649	0.987828	5.35119	New.York	3.1737192	1.3561893	6.357143
Montreal	3.461157	0.9884434	5.702381	Boston	2.7745278	0.8265343	4.660714
Beersheba	2.034785	0.7076567	3.988095	Tel.Aviv.District	2.9124874	0.6686352	6.083333
Eilat	3.6484616	1.0176405	6.184524	Haifa	3.0790735	1.0599728	7.922619
Nahariyya	2.8918947	0.9537472	7.869048	Jerusalem	1.4064533	0.8163763	5.553571

Table 8. Summary statistics for wind direction of the second sample(weekly records)

City	Mean	SD	Median	City	Mean	SD	Median
Vancouver	152.4844	39.63026	212.0714	Portland	193.3868	60.63139	299.3333
San.Francisco	214.1795	43.41145	287.0238	Seattle	180.7629	61.78335	326.8542
Los.Angeles	140.2806	55.93296	229.8512	San.Diego	204.5378	49.45414	278.8452
Las.Vegas	164.8039	23.47239	214.7738	Phoenix	188.1813	62.45558	316.3095
Albuquerque	210.3578	37.75698	281.2083	Denver	162.9796	42.59023	224.9464
San.Antonio	140.656	26.33136	217.625	Dallas	169.7804	28.24975	251.4345
Houston	146.4601	37.13376	227.2143	Kansas.City	184.5819	40.94217	259.4524
Minneapolis	196.0566	43.22311	279.9162	Saint.Louis	187.3942	30.81072	253.491
Chicago	194.2598	43.28406	298.0714	Nashville	177.0295	30.02461	234.1369
Indianapolis	192.4849	33.93313	261.5	Atlanta	189.9282	50.95443	311.0833
Detroit	196.5428	43.24348	262.2635	Jacksonville	130.6037	47.2489	236.25
Charlotte	166.1933	34.96408	229.2275	Miami	137.6806	33.20761	223.4583
Pittsburgh	198.8997	35.10356	250.5625	Toronto	199.3987	36.84956	260.4192
Philadelphia	202.2631	39.31302	271.0903	New.York	203.4948	39.59365	286.1964
Montreal	194.5114	30.43563	242.1607	Boston	188.887	40.85981	262.7725
Beersheba	207.5152	76.36214	295.5595	Tel.Aviv.District	171.951	34.46227	230.5952
Eilat	143.9289	36.43999	241.6905	Haifa	188.1833	72.33378	276.9464
Nahariyya	187.5135	73.23475	283.6548	Jerusalem	199.3459	73.30682	292.494

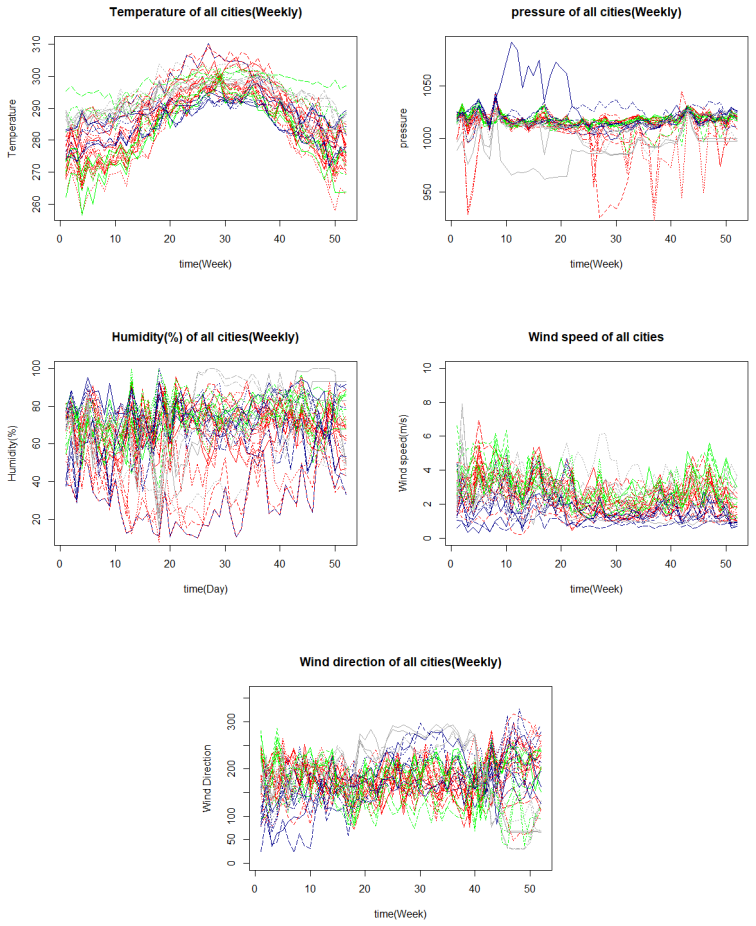


Figure 72. The raw weekly temperature(K) data; the raw weekly air pressure (hPa) data; the raw weekly humidity(%) data; the raw weekly wind speed(m/s) data; the raw weekly wind direction(meteorological degrees). The Atlantic coastal cities in green; pacific coastal cities in blue; the continental cities in red and Mediterranean coastal cities in gray.

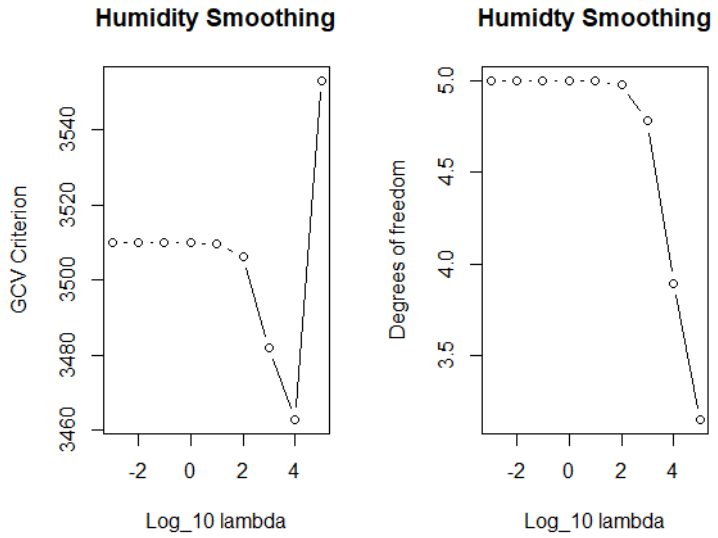


Figure 73. GCV for lambda values $\log_{10}(\lambda) = -2, 0, 2, 4, 6$, for weekly humidity records. The roughness penalty was defined by harmonic acceleration

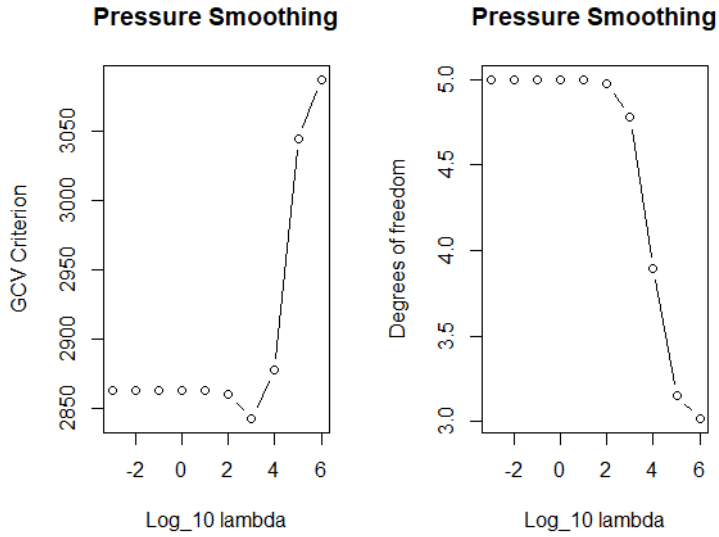


Figure 74. GCV for lambda values $\log_{10}(\lambda) = -2, 0, 2, 4, 6$, for weekly pressure records. The roughness penalty was defined by harmonic acceleration

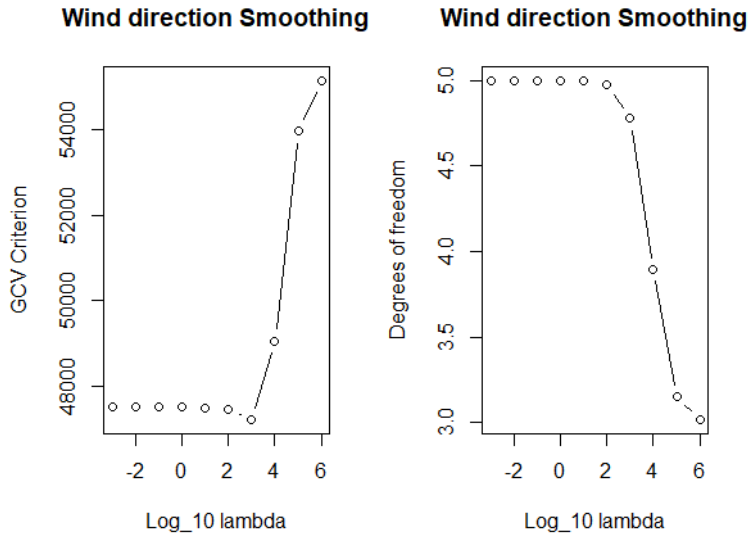


Figure 75. GCV for lambda values $\log_{10}(\lambda) = -2, 0, 2, 4, 6$, for weekly wind direction records. The roughness penalty was defined by harmonic acceleration

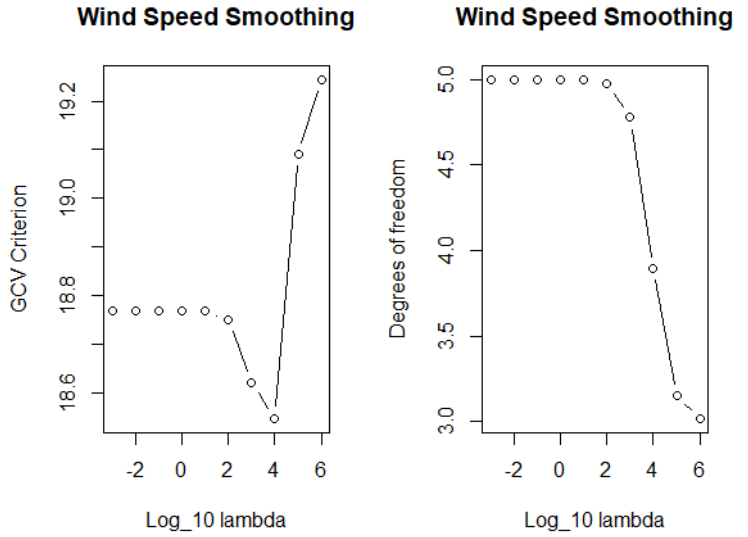


Figure 76. GCV for lambda values $\log_{10}(\lambda) = -2, 0, 2, 4, 6$, for weekly wind speed records. The roughness penalty was defined by harmonic acceleration

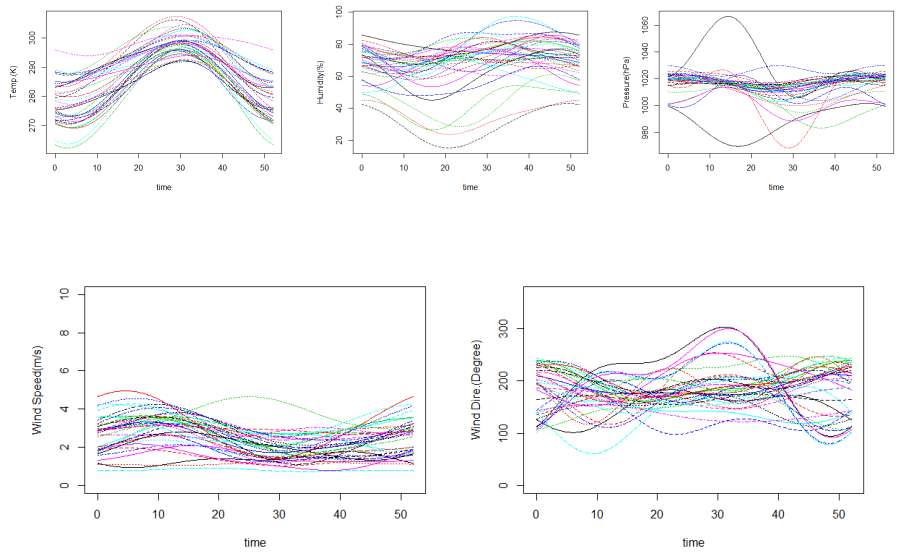


Figure 77. (i) Smoothing temperature curves. (ii) Smoothing humidity curves. (iii) Smoothing pressure curves. (iv) Smoothing wind speed curves. (v) Smoothing wind direction. All curves smoothed by Fourier basis with 5 terms.

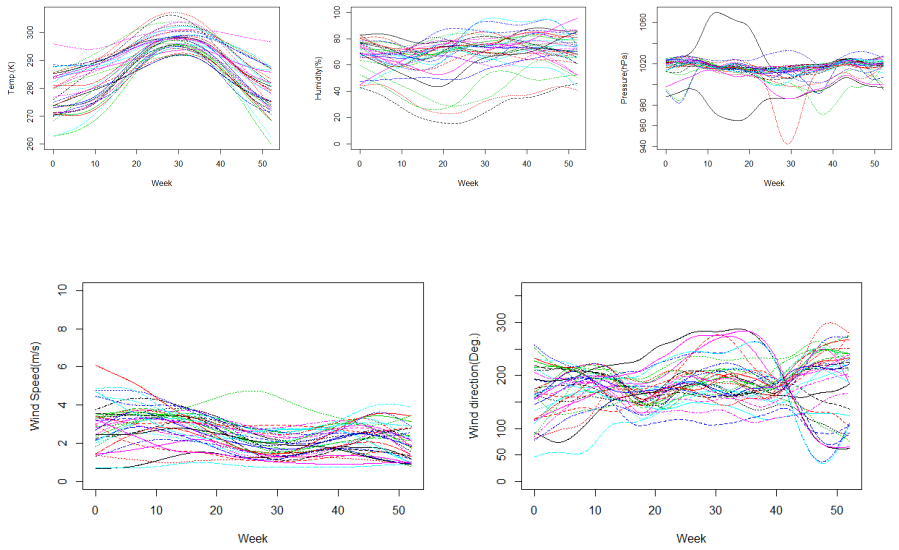


Figure 78. (i) Smoothing temperature curves. (ii) Smoothing humidity curves. (iii) Smoothing pressure curves. (iv) Smoothing wind speed curves. (v) Smoothing wind direction. All curves smoothed by B-spline basis with 4 terms.

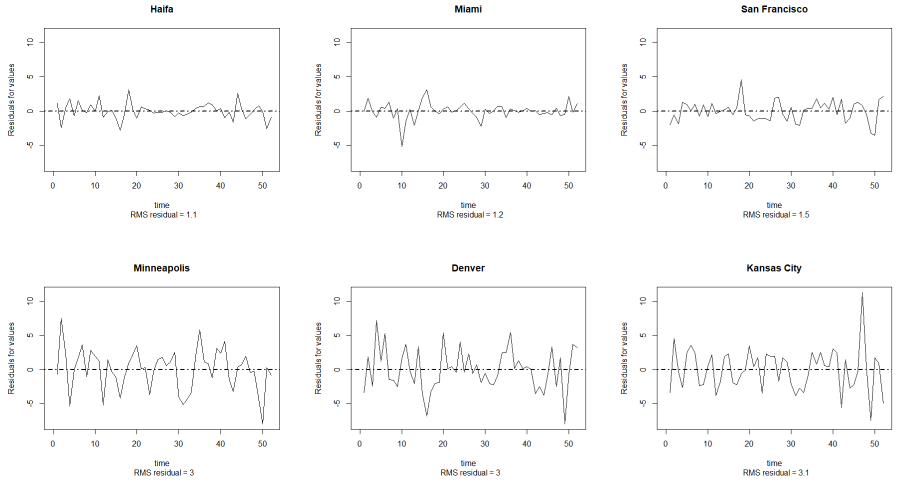


Figure 79. (a)-(c) The residuals for three best fitting curves of temperature using Fourier basis expansion. (d)-(f) The residuals for three worst fitting curves of temperature using smoothing by Fourier basis.

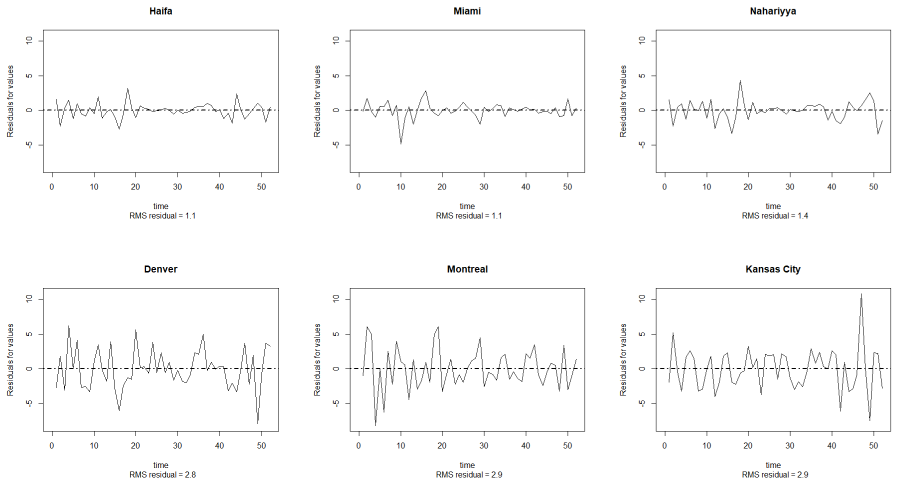


Figure 80. (a)-(c) The residuals for three best fitting curves of temperature using B-splines basis. (d)-(f) The residuals for three worst fitting curves of temperature using smoothing by B-spline basis.

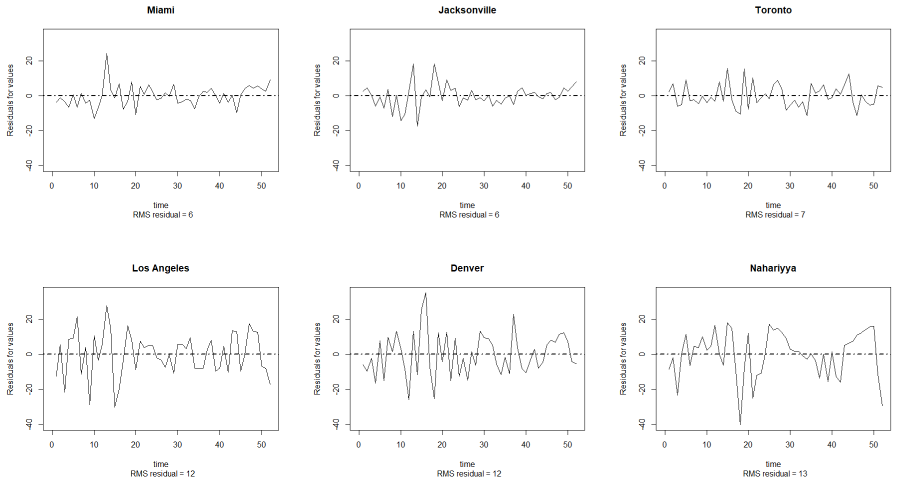


Figure 81. (a)-(c) The residuals for three best fitting curves of humidity using Fourier basis. (d)-(f) The residuals for three worst fitting curves of humidity using smoothing by Fourier basis.

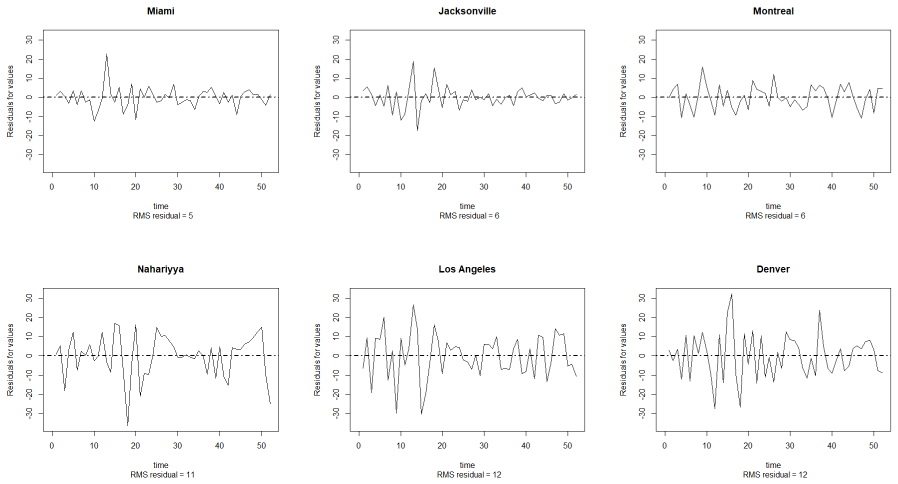


Figure 82. (a)-(c) The residuals for three best fitting curves of humidity using B-spline basis. (d)-(f) The residuals for three worst fitting curves of humidity using smoothing by B-spline basis.

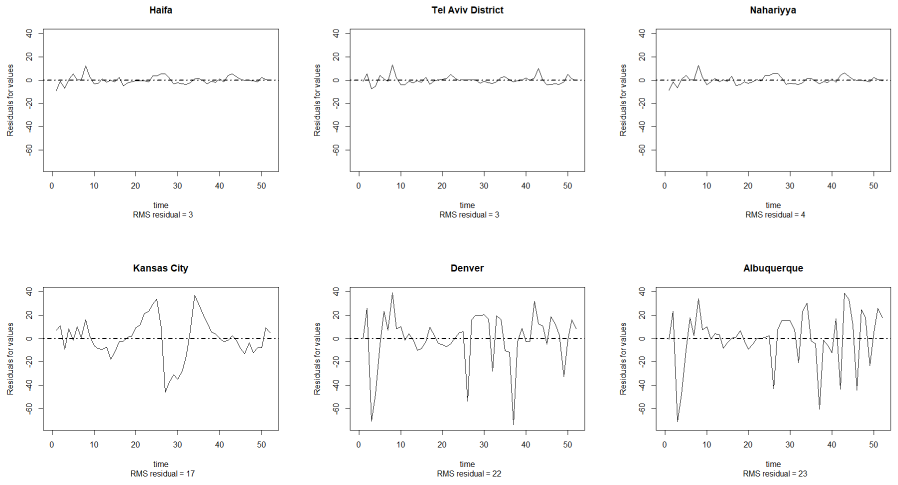


Figure 83. (a)-(c) The residuals for three best fitting curves of pressure using Fourier basis. (d)-(f) The residuals for three worst fitting curves of pressure using smoothing by Fourier basis.

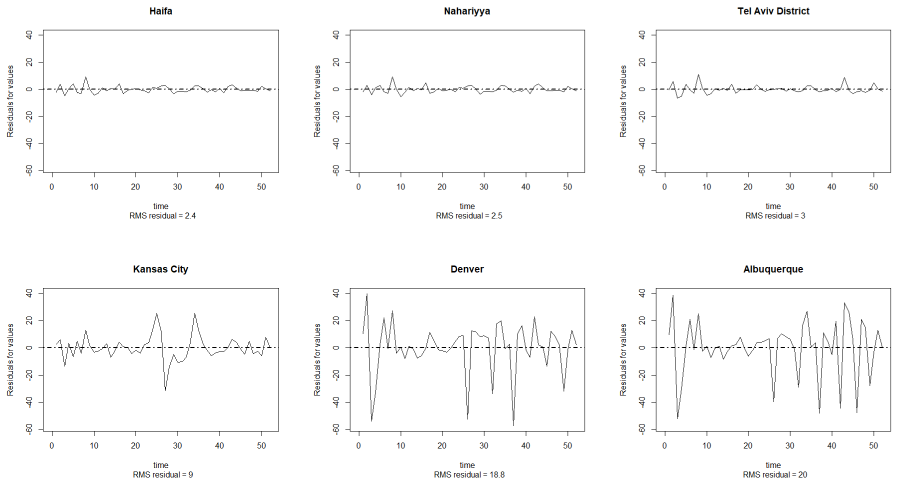


Figure 84. (a)-(c) The residuals for three best fitting curves of pressure using B-spline basis. (d)-(f) The residuals for three worst fitting curves of pressure using smoothing by B-spline basis.

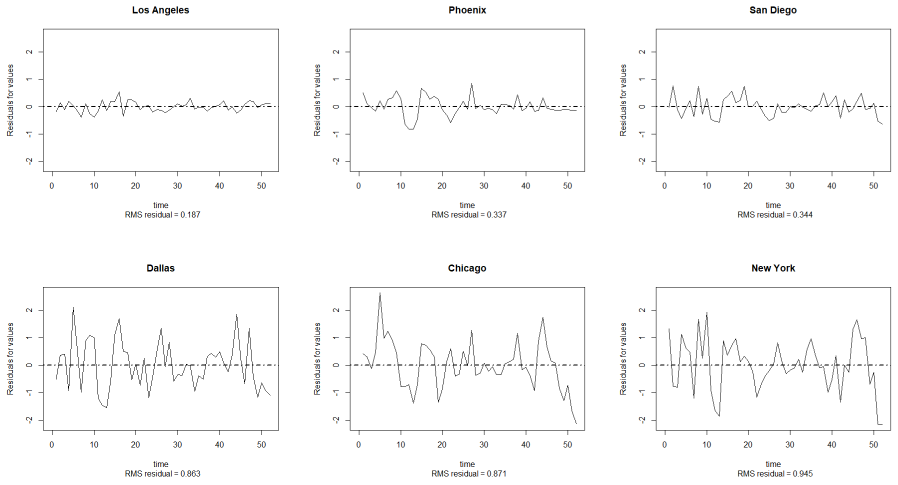


Figure 85. (a)-(c) The residuals for three best fitting curves of wind speed using Fourier basis. (d)-(f) The residuals for three worst fitting curves of wind speed using smoothing by Fourier basis.

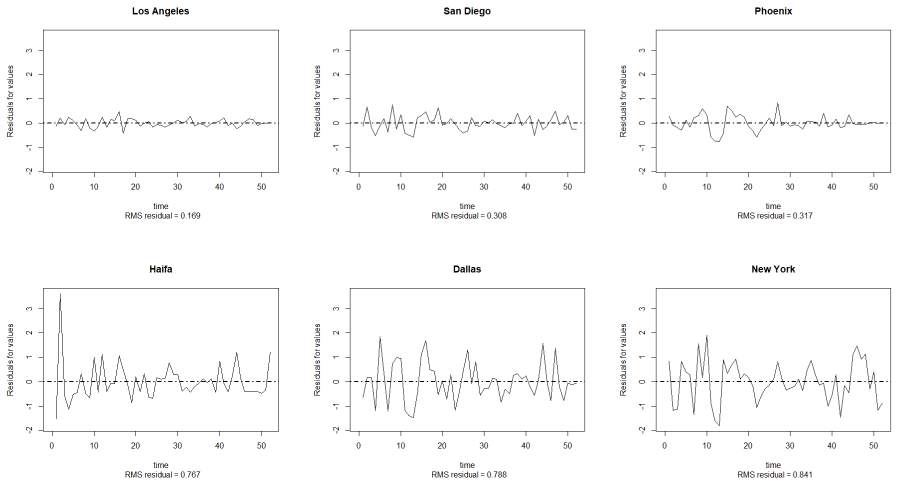


Figure 86. (a)-(c) The residuals for three best fitting curves of wind speed using B-spline basis. (d)-(f) The residuals for three worst fitting curves of wind speed using smoothing by B-spline basis.

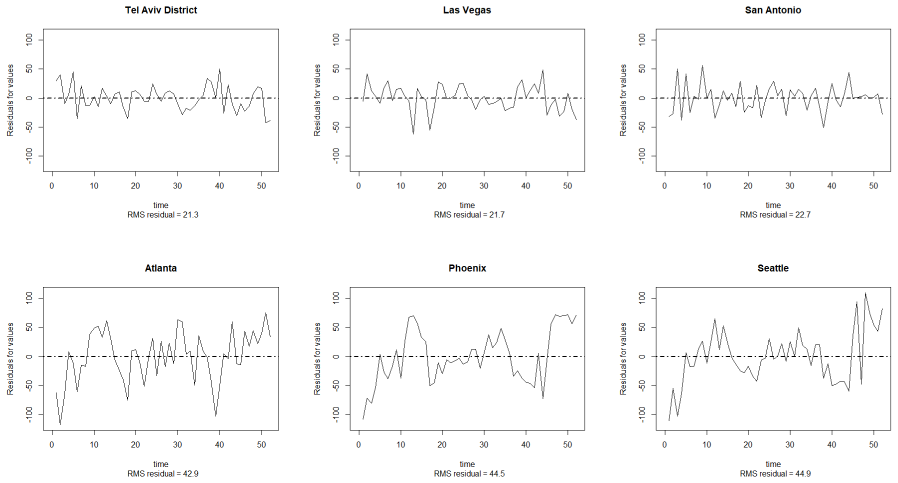


Figure 87. (a)-(c) The residuals for three best fitting curves of wind direction using Fourier basis. (d)-(f) The residuals for three worst fitting curves of wind direction using smoothing by Fourier basis.

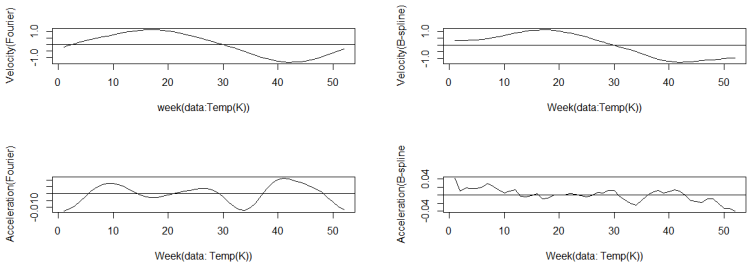


Figure 89. The upper panel shows the first derivative of weekly temperature and lower panel is the second derivative of weekly temperature.

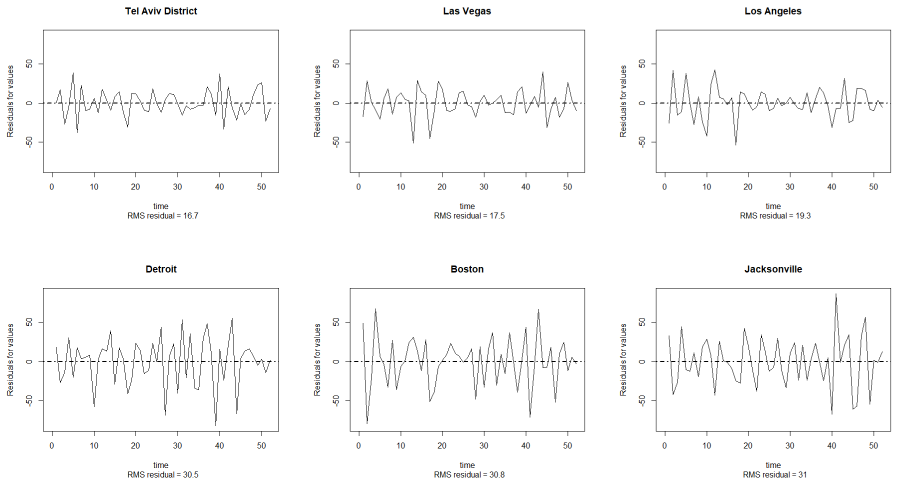


Figure 88. (a)-(c) The residuals for three best fitting curves of wind direction using B-spline basis. (d)-(f) The residuals for three worst fitting curves of wind direction using smoothing by B-spline basis.

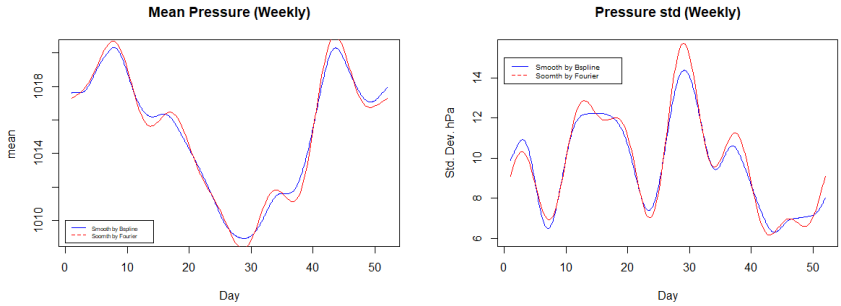
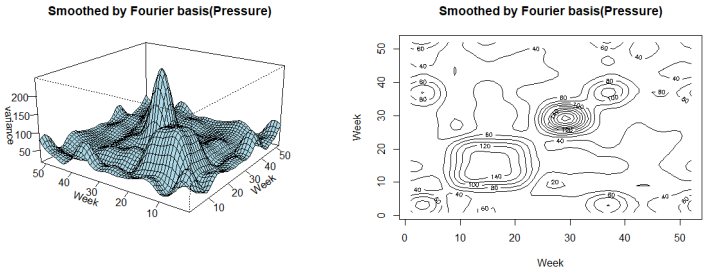
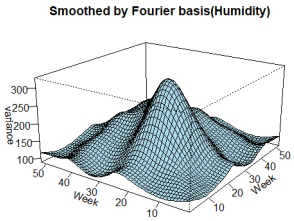


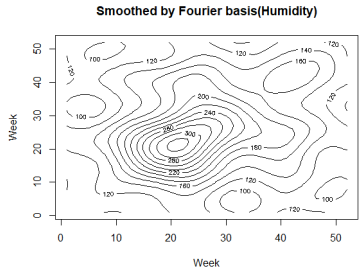
Figure 90. The mean and standard deviation function for smoothed weekly pressure of the 36 cities smoothing by Fourier basis and B-spline.



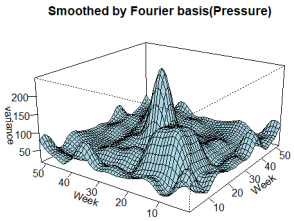
(i) (ii)

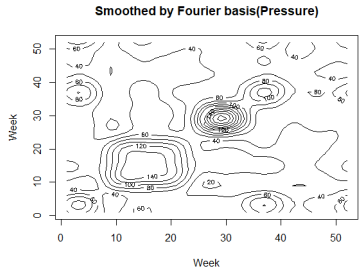
Figure 91. The left panel is a perspective plot of the bivariate correlation function values for the weekly pressure data smoothing by Fourier basis. The right panel shows the same surface by contour plotting. Time is measure in week.



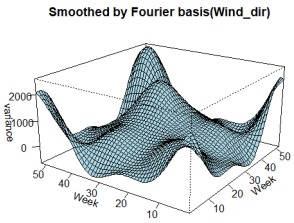


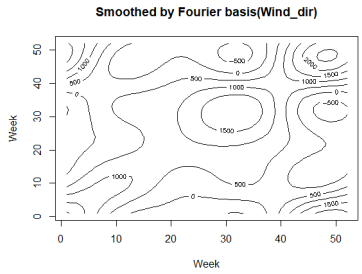
(i) (ii)



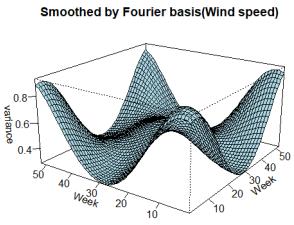


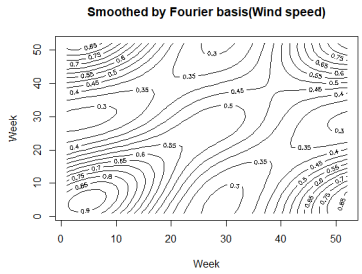
(iii) (iv)





(v) (vi)





(vii) (viii)

Figure 92. Fitted smooth covariance surface at time t.

Table 9. Values of test statistics and p-values of all tests for monthly / weekly / daily humidity(%) data for [a,b]=[0,1]

Test	monthly humidity data		weekly humidity data		daily humidity data	
	Test stat.	p-value	Test stat.	p-value	Test stat.	p-value
CH	60430	0.1368	420544.6	0.054	3405062	0.0259
CS	60430	0.074	420544.6	0.026	3405062	0.0147
L2N	13185.72	0.07527099	86817.17	0.02158229	703230.2	0.005121522
L2B	13185.72	0.06246412	86817.17	0.01566294	703230.2	0.002829495
L2b	13185.72	0.0555	86817.17	0.0171	703230.2	0.0096
FN	2.02817	0.09383962	2.387249	0.0322534	2.51328	0.009425325
FB	2.02817	0.09039243	2.387249	0.02815819	2.51328	0.006517323
Fb	2.02817	0.0865	2.387249	0.0455	2.51328	0.0212
GPF	2.098127	0.07119105	2.354736	0.02329092	2.653862	0.002695601
Fmaxb	4.362853	0.0633	8.177968	0.0104	17.55597	4e-04
TRP	-	1.557621e-10	-	1	-	6.129653e-08
FP	2.034817	0.082	2.384483	0.024	2.511479	0.011

Notes: TRP - tests based on K = 30 random projections and p-value ANOVA without permutation.

Table 10. Values of test statistics and p-values of all tests for monthly / weekly / daily pressure(hPa) data for [a,b]=[0,1]

Test	monthly air pressure		weekly air pressure		daily air pressure	
	Test stat.	p-value	Test stat.	p-value	Test stat.	p-value
CH	56285.75	0.0274	259055.4	0.017	2051566	0.0204
CS	56285.75	0.0312	259055.4	0.0262	2051566	0.0116
L2N	11534.73	0.003616243	51882.85	0.001436771	418900.9	0.004038772
L2B	11534.73	0.002015064	51882.85	0.0006091707	418900.9	0.001919273
L2b	11534.73	0.0286	51882.85	0.0265	418900.9	0.0152
FN	2.755606	0.007284019	2.703271	0.003281729	2.374403	0.007493818
FB	2.755606	0.005205228	2.703271	0.001840787	2.374403	0.00447227
Fb	2.755606	0.0412	2.703271	0.0425	2.374403	0.0355
GPF	3.042881	0.001445816	3.874619	4.265625e-06	4.112353	1.046088e-08
Fmaxb	6.432095	0.011	12.78454	0.0013	24.51136	0
TRP	-	2.906637e-06	-	0.001389225	-	1
FP	2.773687	0.012	2.700018	0.002	2.374552	0.007

Notes: TRP - tests based on K = 30 random projections and p-value ANOVA without permutation.

Table 11. Values of test statistics and p-values of all tests for monthly / weekly / daily wind speed(m/s) data for [a,b]=[0,1]

Test	monthly wind speed		weekly wind speed		daily wind speed	
	Test stat.	p-value	Test stat.	p-value	Test stat.	p-value
CH	462.0966	0.0063	2598.426	0.0015	29875.31	0
CS	462.0966	0.0025	2598.426	0.0014	29875.31	6e-04
L2N	108.7967	4.590971e-07	586.1731	3.094009e-09	6454.015	2.187917e-12
L2B	108.7967	9.770138e-08	586.1731	1.033963e-10	6454.015	0
L2b	108.7967	0	586.1731	2e-04	6454.015	2e-04
FN	5.804739	1.244521e-05	5.08179	2.052178e-07	3.918484	3.267457e-10
FB	5.804739	5.564039e-06	5.08179	2.041599e-08	3.918484	2.498002e-14
Fb	5.804739	0.0011	5.08179	6e-04	3.918484	8e-04
GPF	5.523986	1.09736e-06	4.906053	1.286082e-08	3.829353	2.401257e-11
Fmaxb	9.157585	0.0022	12.30951	7e-04	22.48578	1e-04
TRP	-	1	-	1	-	1
FP	5.855598	0	5.083232	0	3.920896	0

Notes: TRP - tests based on K = 30 random projections and p-value ANOVA without permutation.

Table 12. Values of test statistics and p-values of all tests for monthly / weekly / daily wind direction(meteorological degrees) data for [a,b]=[0,1]

Test	monthly wind direction		weekly wind direction		daily wind direction	
	Test stat.	p-value	Test stat.	p-value	Test stat.	p-value
CH	1824698	0	9316409	0	94760791	0
CS	1824698	0	9316409	0	94760791	0
L2N	344440.7	3.654743e-12	1785191	5.884182e-15	18432611	0
L2B	344440.7	1.679767e-13	1785191	0	18432611	0
L2b	344440.7	0	1785191	0	18432611	0
FN	8.747375	9.593759e-09	7.065632	3.915113e-11	4.812142	6.838974e-14
FB	8.747375	1.957259e-09	7.065632	6.568079e-13	4.812142	0
Fb	8.747375	0	7.065632	1e-04	4.812142	0
GPF	7.568427	3.350107e-10	6.479825	1.476597e-14	4.953651	0
Fmaxb	12.51006	1e-04	14.06415	9e-04	15.17025	0.0016
TRP	-	1.561018e-10	-	5.291812e-08	-	4.863443e-11
FP	8.704647	0	7.097725	0	4.818358	0

Notes: TRP - tests based on K = 30 random projections and p-value ANOVA without permutation.