UNIVERSITI PUTRA MALAYSIA

FUNCTIONAL EXTREME DATA ANALYSIS METHODS AND ITS APPLICATION TO RAINFALL DATA

NOOR IZYAN BINTI MOHAMAD ADNAN

IPM 2018 11

FUNCTIONAL EXTREME DATA ANALYSIS METHODS AND ITS APPLICATION TO RAINFALL DATA

By

NOOR IZYAN BINTI MOHAMAD ADNAN

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfillment of the Requirements for the Doctor of Philosophy

January 2018

DEDICATIONS

To my parent, Mohamad Adnan and Nurizan, for their steadfastness in prayer.
To my parent-in-law, Ton Mohamed and Wan Hazanah, for their affection.
To my husband, Adie Safian, for his Love and understanding.
To my sons, Adam Saufi and Ammar Shauqi: You are such a brilliant princes!
To my sisters, brothers, and in-laws, for their support.

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfillment of the requirement for the degree of Doctor of Philosophy

FUNCTIONAL EXTREME DATA ANALYSIS METHODS AND ITS APPLICATION TO RAINFALL DATA

By

NOOR IZYAN BINTI MOHAMAD ADNAN

January 2018

Chairman: Mohd Bakri Adam, PhD
Institute: Institute for Mathematical Research

Functional data analysis is one of the new techniques to transform a discrete or continuous observation into a functional form. Conventional classical statistics methods now can be observed and analyzed through a curve for almost all types of data with neither distribution assumption nor goodness of fit test that are necessary to be followed. Literature reviews show that there is no study found for functional data analysis application on extreme data which deals with maximum value in the data set.

In this thesis, the study has extended the functional data analysis methodology to cover on extreme data with several substitution methods have been introduced. Some characteristics of functional extreme data analysis such as on environmental data are explained. The tolerance bands for functional mean extreme data is proposed using bootstrapping method by implementing the percentile computation in determining the upper and lower limits of the mean function.

The performance of the functional extreme data analysis is carried out. The equal and unequal space of time cases are considered to be implemented for the functional extreme data. The study found that only data that consists a large number of extreme data will be performed in functional extreme data for unequal space of time. Otherwise, a small number of extreme data is suggested to use the equal space of time to obtain a smooth curve.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

ANALISIS FUNGSIAN BAGI DATA EKSTRIM DAN APLIKASINYA TERHADAP DATA TABURAN HUJAN

Oleh

NOOR IZYAN BINTI MOHAMAD ADNAN

Januari 2018

Pengerusi: Mohd Bakri Adam, PhD
Institut: Institut Penyelidikan Matematik

Analisis data fungsian adalah salah satu kaedah baru yang digunakan bagi mengubah data berbentuk diskret atau berterusan kepada data yang berbentuk fungsian. Kaedah statistik yang klasik kini telah dapat dianalisa menerusi lengkuk bagi hampir keseluruhan jenis data dengan tiada andaian taburan atau ujian penyuaian model diperlukan. Kajian literatur menunjukkan masin tiada kajian yang menggaplikasikan analisis data fungsian terhadap data ekstrim yang melibatkan nilai maksima di dalam sesebuah set data.

Kaedah analisa data fungsian ini telah dipanjangkan aplikasinya terhadap data ekstrim dan beberapa kaedah baru juga telah di perkenalkan di dalam kajian ini. Ciri-ciri tertentu yang dimiliki oleh analisis data fungsian contohnya terhadap data persekitaran juga dibincang dan dijelaskan dengan terperinci. Had toleransi bagi data ekstrim dicadangkan dengan menggunakan kaedah bootstrapping melalui pengiraan peratusan yang diterapkan didalam mencari had atas dan bawah purata sesuatu fungsi.

Kemudian, ujian tahap prestasi analisis data ekstrim fungsian dijalankan. Situasi samada jarak atau selang masa yang sama atau berbeza untuk diaplikasikan terhadap data ekstrim fungsian juga diambil kira. Kajian mendapati bahawa hanya data ekstrim yang besar sahaja yang sesuai bagi selang masa yang berbeza. Manakala, bagi data ekstrim yang mempunyai bilangan data yang kecil adalah dicadang untuk menggunakan selang masa yang sama bagi menghasilkan lengkuk yang licin.

ACKNOWLEDGEMENTS

Praise be to Allah the Almighty and Merciful, Who has given me an enormous miracle in every struggle, so that I can finish my thesis entitled ***"Functional Extreme Environmental Data Analysis"***. Peace upon the prophet Muhammad S.A.W who has brought Islamic norms and values to the entire world.

I would like to express my profound gratitude to Associate Prof. Dr. Mohd Bakri Adam who giving me the chance to work with him whilst guiding my first attempt to in depth scientific research. I am also grateful to Prof. Dr. Noor Akma Ibrahim, Dr. Mohd Yusoff Ishak and Dr. Mohammad Noor Amal Azmai, member of my supervisory committee, for their useful comments.

I also acknowledge the staff of the Institute for Mathematical Research (INSPEM) for providing a friendly working environment. I am grateful to my fellow students under the supervision of Associate Prof. Dr. Mohd Bakri Adam. I appreciate my fellow INSPEM students whom we share the ideas and moment together.

Special thanks to my parents, my siblings and my in-laws for their prayers and support have kept me during my study.

Most of all, I wish to thank my husband, Adie Safian and my sons, Adam Saufi and Ammar Shauqi for without their support and patience this work mean nothing!

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Doctor of Philosophy. The members of the Supervisory Committee were as follows:

Mohd Bakri Adam, PhD
Assosiate Professor
Institute for Mathematical Research
Universiti Putra Malaysia
(Chairperson)

Noor Akma Ibrahim, PhD
Professor
Institute for Mathematical Research
Universiti Putra Malaysia
(Member)

Mohd Yusoff Ishak, PhD
Senior Lecturer
Faculty of Environmental Studies
Universiti Putra Malaysia
(Member)

Mohammad Noor Amal Azmai,
Senior Lecturer
Faculty of Science
Universiti Putra Malaysia
(Member)

ROBIAH BINTI YUNUS, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

Declaration by graduate student

I hereby confirm that:

- this thesis is my original work;
- quotations, illustrations and citations have been duly referenced;
- this thesis has not been submitted previously or concurrently for any other degree at any other institutions;
- intellectual property from the thesis and copyright of thesis are fully-owned by Universiti Putra Malaysia, as according to the Universiti Putra Malaysia (Research) Rules 2012;
- written permission must be obtained from supervisor and the office of Deputy Vice-Chancellor (Research and Innovation) before thesis is published (in the form of written, printed or in electronic form) including books, journals, modules, proceedings, popular writings, seminar papers, manuscripts, posters, reports, lecture notes, learning modules or any other materials as stated in the Universiti Putra Malaysia (Research) Rules 2012;
- there is no plagiarism or data falsification/fabrication in the thesis, and scholarly integrity is upheld as according to the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) and the Universiti Putra Malaysia (Research) Rules 2012. The thesis has undergone plagiarism detection software.

Signature:____________________________Date:____________________________

Name and Matric No: Noor Izyan Binti Mohamad Adnan

Declaration by Members of Supervisory Committee

This is to confirm that:

- the research conducted and the writing of this thesis was under our supervision;
- supervision responsibilities as stated in the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) are adhered to.

Signature: ______________________________
Name of
Chairman of
Supervisory
Committee: Mohd Bakri bin Adam

Signature: ______________________________
Name of
Member of
Supervisory
Committee: Noor Akma binti Ibrahim

Signature: ______________________________
Name of
Member of
Supervisory
Committee: Mohd Yusoff bin Ishak

Signature: ______________________________
Name of
Member of
Supervisory
Committee: Mohammad Noor Amal bin Azmai

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure **Page**

LIST OF ABBREVIATIONS

AIC	Akaike Information Criteria
AICc	Corrected Akaike Information Criteria
B	Number of re-samples
CV	Coefficient of Variation
EB	Extra Balance
EVT	Extreme Value Theory
FANOVA	Functional Analysis of Variance
FDA	Functional Data Analysis
FEDA	Functional Extreme Data Analysis
FPC	Functional Principal Component
GCV	Generalized Cross-Validation
GEV	Generalised Extreme Value
GPD	Generalised Pareto Distribution
IDW	Inverse Distance Weight
K	Number of Basis Function
LP	Lower Point
m	Number of block
MRL	Mean Residual Life
n	Number of observations
N	Number of Functional Data
NR	Normal Ratio
NRIDW	Normal Ratio with Inverse Distance Weight
PENSSE	Penalized Sum Square Error
POT	Peak Over Threshold
pdf	Probability Density Function
r	Most highest values in each block
SI	Smoothing Index
SIR	Smoothing Index Rank
SSE	Sum Square Error
t	Time
T	Number of time
UP	Upper Point

CHAPTER 1

INTRODUCTION

This chapter presents the theory of the functional data analysis and extreme value data with some related references in the functional data along with an extreme value application over the various field of study. Moreover, the motivation of the study, objectives of the study, expected outcome, scope of the study and significance of the study as well as the structure of the thesis are also described here.

1.1 Background

”Functional Data Analysis” is known as FDA and an extreme data are two important things to be comprehended in this study. Method of functional data analysis will be used as a main technique to the extreme environmental data. New approach of functional data analysis might give a different point of view to the data with a functional observations as a replacement of discrete observations. At the same time, the extreme value data will possibly be implemented by considering only the highest values present in the data. The development of FDA methods can be seen through various problems occurred in many fields of study, Ramsay and Silverman (2002). Hence, this study proposes a method development of functional data analysis for extreme data and the application of FDA on the extreme environmental data. The entire all three extreme value data approaches are tested to FDA which can be listed as Block Maxima, *r*-Largest Order Statistics, and Peak Over Threshold (POT). The extreme data used for this study will be transformed into functional form. The idea of how FDA and extreme data work are explained in the next subsection.

1.1.1 Functional Data Analysis

One Function

Either raw discrete or continuous data which is widely used as a current conventional approach in various research areas are easy to be implemented and dealt with as numerous data are collected and presented in this form. A discrete or continuous observations of one data set values always be represented as $\{x_1, x_2, x_3, \ldots, x_T\}$ or simply written as x_t for data which are recorded in time, t for $t = 1, \ldots, T$ where T is the finite value of t. This conventional method gives less information especially for an illustration of the pattern to shows the behavior of the collected data.

As can be seen in Figure 1.1, an **example** of plotting a raw continuous data as a single observation over the time for $t = 1, \ldots, 36$. The plot shows that it is not enough evidence to obtain towards the behavior of the data. Nothing much can be achieved rather than a summary value of the data like a minimum, maximum, range, mean and median value. The most general behavior that can be seen here there is an increasing pattern along the time.

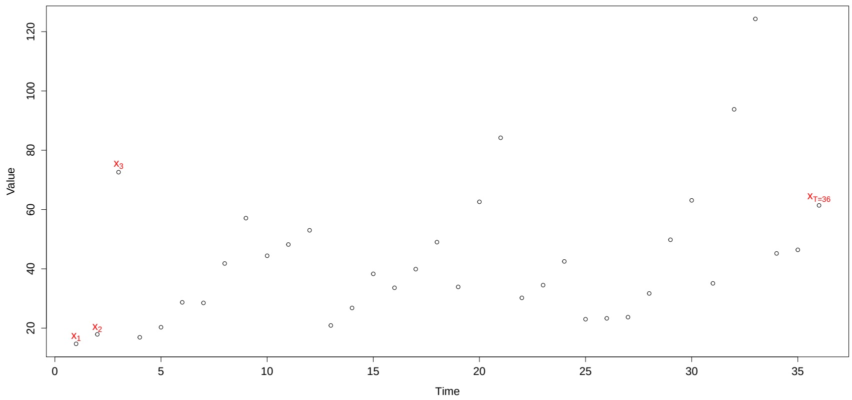

Figure 1.1: Example of plotting the 36 raw continuous data over time. Each point treated as a single observation.

One can easily use time series analysis in obtaining the pattern of the data which also provide forecast for the future pattern. Unfortunately, time series analysis only work for equally time space and required for stationary assumption. However, compared to the functional data analysis, FDA provides additional and extensive evidence to the research that aid further analysis without any restriction where the data can be arranged in unequal time space and no assumptions need to be followed.

Influenced by the features of FDA, the data used in this study is in functional form. FDA basically represents the observations x_t from the raw data into a functional form or functional observation. The number of x_t which is represent for the observations of raw data is denoted as n. FDA will represents the entire x_t as one function denoted as $f(t)$. This functional form of function $f(t)$ contains the fitted values of f_t that are also recorded in time t, which can be written as $f(t) = \{f_1, f_2, f_3, \ldots, f_T\}$. The process of how to transform the raw of discrete or continuous data into functional form will be discussed in Chapter 3.

Figure 1.2 shows the **example** of smooth curve of functional observation of function $f(t)$ for the fitted values of f_t for $t = 1, \ldots, 36$. The raw data of x_t known as observed values represent by white points while the fitted values of f_t represent by the black points. It is obviously illustrates that the raw data from the conventional approach is actually has an interesting pattern to be explored which need an advance technique such as FDA.

There is a situation where each value of observed data x_t, are equal to each fitted values f_t, where $x_t = f_t$. This is known as interpolation, where the data is trusted with no error presents, that need to be removed. This type of function can be seen from Figure 1.3 where all the points are joint to obtain one functional form.

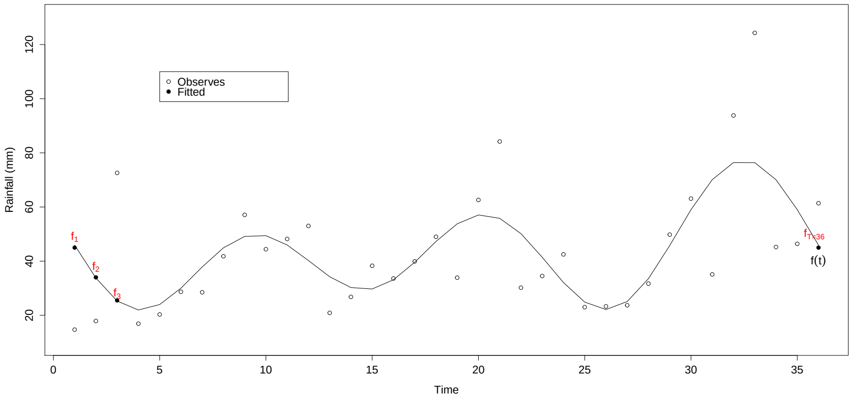

Figure 1.2: Example of smooth functional observation represented by function $f(t)$. The white points are the observed values or raw data x_t, the black points are the fitted values f_t.

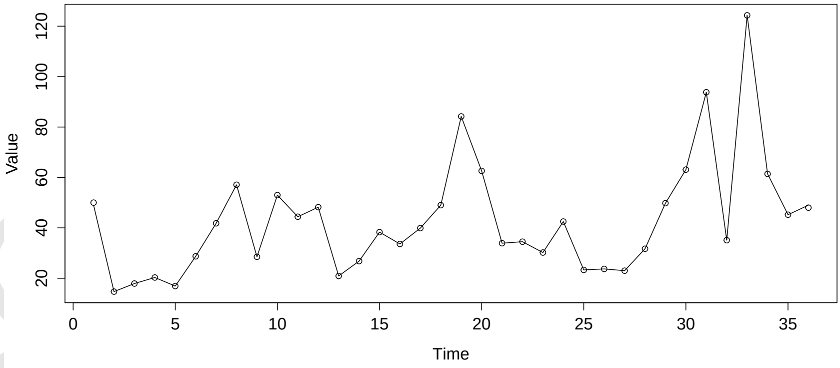

Figure 1.3: Example of interpolation of functional observation. The raw data have the same value with the fitted values, $x_t = f_t$.

For a large data, taking every value at time, t would be loaded for a huge number of values, Ramsay and Silverman (2005). Thus, a smooth function is advised so that the derivatives of the function can be obtained which enhance an excellent features of the data.

Therefore, there are subtle features in functional data which are not appear and available in a raw discrete and continuous data. Besides, FDA also ables to analyze and examine what the raw discrete and continuous data from classical statistics can do. FDA not only deal with a single (short or long) record, nevertheless FDA also permits independent replications, Ramsay and Silverman (2005). This types of functional data explained in the next subsection.

Many Functions

Functional data in general consists of replication and regularity as a features. The replication is refer to the number of subject in the data. Replication is a summarization across the curves and regularity involves an exploration of information within the curves, Ramsay and Silverman (2005).

Instead of having only one function or one subject, the function can be replicated to i subject for $i = 1, \dots, N$, where N is the number of functional observations. Then the i^{th} function can be written as $f_i(t)$ with each i treated as a different subject. The raw data values for replication i can be written as $\{x_{i,1}, x_{i,2}, x_{i,3}, \dots, x_{i,T}\}$, and the function $f_i(t)$ has a list of converted values as $\{f_{i,1}, f_{i,2}, f_{i,3}, \dots, f_{i,T}\}$ over the time t .

Figure 1.4 depicts an **example** of one replication functional observation together with the fitted values along the function for $t = 1, \dots, 36$. The curve is denoted as $f_1(t)$ represents for the first functional observation. The values for this first function denoted as $f_{1,1}$, $f_{1,2}$, $f_{1,3}$ for the values recorded at $t = 1$, $t = 2$ and $t = 3$ in the function respectively or written as $f_1(t) = \{f_{1,1}, f_{1,2}, f_{1,3}\}$. Whereas $f_{1,T}$ is the fitted value recorded at the last number of t in the first function, as in this example, $f_{1,T} = f_{1,36}$. In general, the i replication of functions can be expressed as $f_i(t) = f_1(t), \dots, f_N(t)$.

Figure 1.5 illustrates **example** of the functional observations with i replication for $N = 5$ subjects. There are functions representing five subjects by $f_1(t)$, $f_2(t)$, $f_3(t)$, $f_4(t)$, $f_5(t)$ with blue, black, turquoise, green and red curves respectively. The replication of the curves from N number of subject propose and discover for the variation occurs between or across the curves.

A new descriptive statistics for FDA towards analysis of variation found by Keser et al. (2016) give a value added in FDA. This different approach in analyzing variation, named as functional coefficient of variation will be completely defined in Chapter 3.

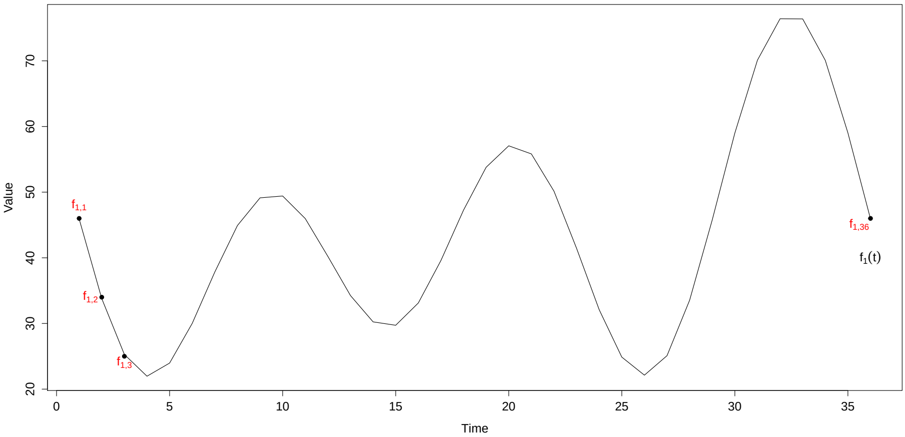

Figure 1.4: Example of one replication of functional observations, $f_1(t)$ with the values of the function, $\{f_{1,1}, f_{1,2}, f_{1,3}, \ldots, f_{1,36}\}$.

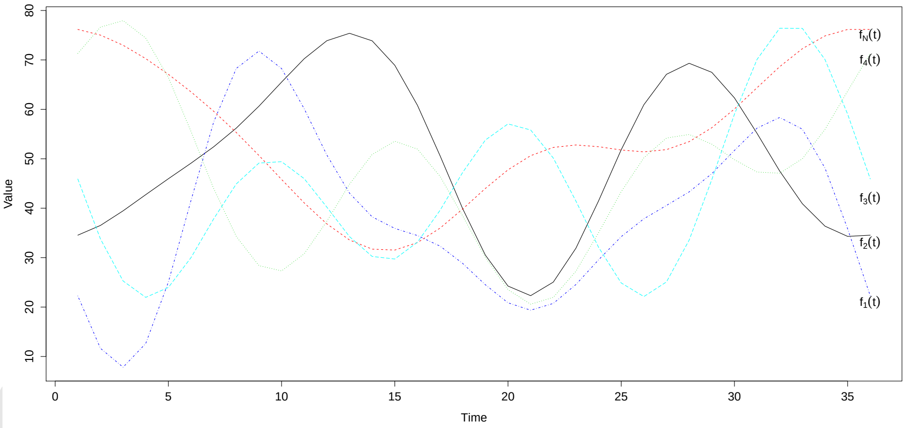

Figure 1.5: Example of more than one functional observations. The first function represents the first subject by $f_1(t)$ as a blue curve. The second, third and fourth subject represent by $f_2(t)$, $f_3(t)$, and $f_4(t)$ as a black, turquoise and green curve respectively. The red curve be the last function that we denoted as $f_N(t)$ represents by $f_5(t)$.

Perspectives on Functional Data Analysis

Functional data analysis able to convert any types of data into functional form and tremendously analyzed these data in a functional way even though each data has a specialty in a certain field in classical statistics.

Functional data analysis aims similar purpose as classical statistics as clarified by Ramsay and Silverman (2005) such as representing data to the advance analysis, displaying the several features of data, indicating the significant cause of pattern and variation among data, and describing variation of the dependent variable through the evidence of the independent variable.

Nevertheless, FDA offers a convenient technique in data arrangement and outcome where FDA does not require for stationary assumption and not limited to the equally time space sample of data as applied in time series analysis. This means that the interval of t which denoted as τ, can be unequal for $t = 1$, $t = 2$, $t = 3$ and so forth. The recorded time also can vary from one subject to another subject.

Functional data is similar to multivariate data in terms of infinite dimension. This takes some difficulties in theory and computation although it contains full of information, Wang et al. (2015). At this point, FDA modelling and analysis require a reduction in dimension and smoothing assist as an instrument of standardization, Wang et al. (2015).

The finest occurrence in FDA is an easy way in performing measurement error since each subject observes for one repeated measurements, and the time recorded in unequally space for every subjects, Wang et al. (2015). On the other hand, FDA gives challenge that caused by the small covariance operator in functional regression and functional correlation measures which encourage the inverse operators to be limitless, Wang et al. (2015).

Functional data will take a few methods or processes to convert a raw data into a functional data. Ramsay and Silverman (2005) explain the details about the transforming process and all methods regarding to the functional data analysis. This process are also easy to be practiced as the functional data analysis is unrestricted by any distribution assumption which is no specific assumptions need to be followed and no goodness of fit test is needed, Ramsay and Silverman (2005).

Basically, an essential assumption of FDA is smoothness, however noisier and less frequent data also permitted, Ramsay and Silverman (2002). Through this functional data, the first view easily detected by eyes are the characteristics and the pattern of the data either increasing, decreasing or fluctuate.

As a preliminary step, the first and second derivatives of the function yield the changes rate and the acceleration of the data respectively. Furthermore, the additional exploration in FDA will create an excessive outcome. The research emphasis the first generation of FDA that deal with the univariate data cases. Wang et al. (2015) classify a complex data objects, multivariate data, correlated data comprises of images or shapes as the next generation in FDA.

The key assumption of FDA is the smoothness of the data where this is to distinguishes the FDA from the multivariate analysis. The smoothness will take into account for an error or noise and the concept of standard regression analysis model is implied.

Consequently as motivated by the error model from a linear regression model, the raw data can be fitted using the functional error model that can be written in a general form as $x = f + e$ with x represents for the observed values from the raw data, f represents the fitted values of smooth function $f(t)$, which are in interval of time, denoted as τ while e is the error term, assumed to be independent and identically distributed with mean zero and variance is constant, Ramsay and Silverman (2005).

Figure 1.6 shows **example** of the model in graphically for $t = 1, \ldots, 36$. The observed value, the fitted value and the value for an error are indicated as x_t, f_t and e_t respectively at a specific time t. As a result, one can conclude that, the functional error is $x_t = f_t + e_t$, where $\hat{x}_t = \hat{f}_t$ and error can be obtained by $e_t = x_t - \hat{f}_t$ at a respective time t.

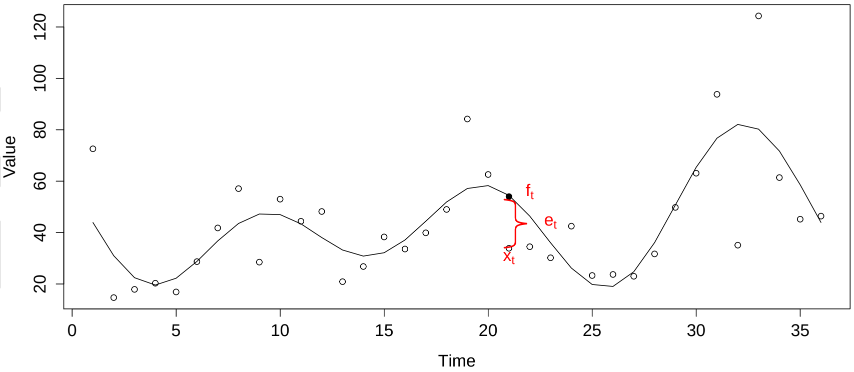

Figure 1.6: The error model of the smooth function $f(t)$. The observed value denoted as x_t, the fitted value denoted as f_t and the error is represented by e_t.

Functional Data Analysis Process

The functional data analysis flow process starts from the beginning of the raw observation to the end of the smoothing function techniques represent in Figure 1.7. The R and S+ software for FDA are available in Ramsay et al. (2009) and Douglas et al. (2005). As an important step, the raw discrete or continuous data have to be transformed into the functional data. According to Ramsay and Silverman (2005), there are two methods of transforming either by interpolation or by smoothing.

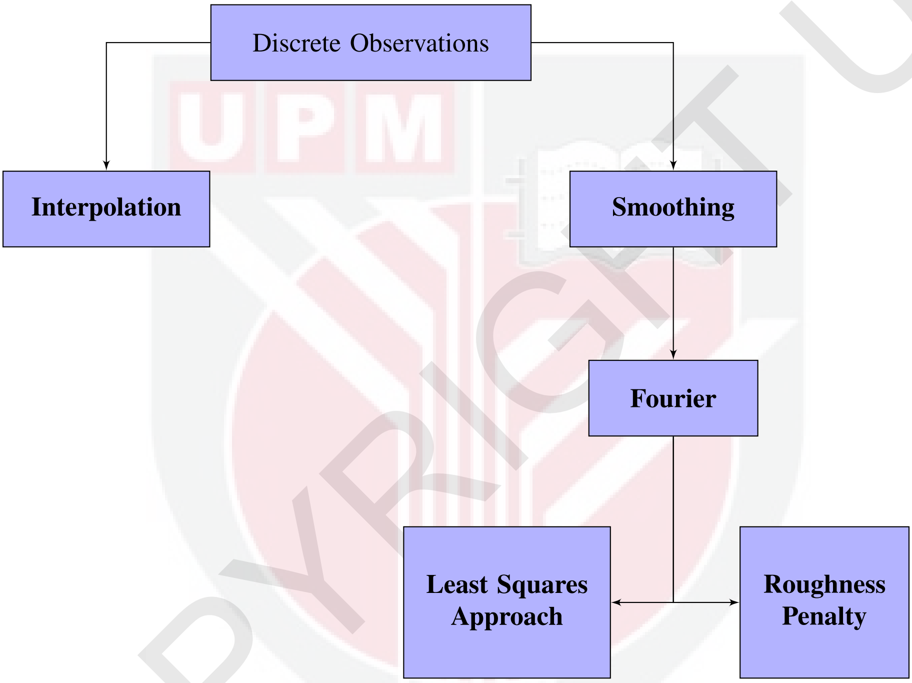

Figure 1.7: Flow diagram of functional data analysis process from raw observation to functional observation. (Ramsay and Silverman, 2005)

The smoothing technique used if the data required an error to take into account in order to obtain a smooth function, Ramsay and Silverman (2005) and Ramsay et al. (2009). While Ramsay and Silverman (2005) state that interpolation scheme which simply connecting all observations using straight line segment might be not an adequate for derivative information, compared to the more advance extraordinary feature such as smoothing technique.

This study considers a smoothing method in a way to represent a non-parametric continuous-time functions by basis expansion methods, where the function $f(t)$ can be examined by

$$f(t) = \sum_{k=1}^{K} \beta_k \phi_k \tag{1.1.1}$$

where K is the maximum number of basis function for $k = 1, \dots, K$. The coefficients of basis function is denoted by β_k and ϕ_k is the basis function. The basis function ϕ_k will be followed a basis system which have several series of functions used to represent the functional data. The familiar series of basis system function can be listed as fourier basis, spline basis, constant basis and monomial basis systems. Ramsay et al. (2009) focus more on fourier and spline basis systems because these systems are complementary of constant and monomial basis systems.

Basically, the fourier basis is usually used for periodic functions which refer to the time series data, while the spline basis intended for non-periodic functional data, Ramsay and Silverman (2005) and Ramsay et al. (2009). There are lots of other potential basis systems such as exponential, polygonal, polynomial and step-function, but those basis systems seem to be less important to functional data analysis, Ramsay et al. (2009). Fourier basis system is selected as a basis function since the study applies a periodic data to be implemented.

A least squares criterion is used in determine the basis coefficient, β_k by minimizing the sum of squared residuals (SSE) which can be written as

$$SSE = \sum_{t=1}^{T} \left[x_t - \sum_{k=1}^{K} \beta_k \phi_k(t) \right]^2, \tag{1.1.2}$$

where $\sum_{k=1}^{K} \beta_k \phi_k(t)$ represent the function $f(t)$ from Equation (1.1.1). The Equation (1.1.2) can be simplifies as

$$SSE = \sum_{t=1}^{T} [x_t - f_t]^2 \tag{1.1.3}$$

where x_t is the observed value and f_t is the fitted value from the functional observation $f(t)$. The least squares approach has a normality assumption for the model $x_t = f_t + \varepsilon_t$ where the residuals ε_t are assumed to be independently and identically distributed with zero mean and variance σ^2 must be constant, Ramsay and Silverman (2005).

In order to prepare the basis function for each basis system, it is necessary to determine the maximum size of the basis required, denoted as K. This number of basis, K will fit the smooth function well. Ramsay and Silverman (2005) and

Ramsay et al. (2009) indicate two ways in defining the number of basis which is through the Least Squares approach and Roughness Penalty process.

The Least Squares is the simplest approach to obtain a small value of K which is K is less than T, whereas the following powerful process of Roughness Penalty will be conducted to compute a large number of K for K is greater than T, Ramsay and Silverman (2005) and Ramsay et al. (2009).

The simplest and straightforward way in determining the number of basis, K is by calculating the unbiased estimated of residual variance, s^2 through the formula defined by

$$s^2 = \frac{1}{T-K}\sum_{t=1}^{T}[x_t - f_t]^2 \tag{1.1.4}$$

where s^2 is the unbiased estimated value of residual variance, σ^2. The maximum value of K then is determined by adding the value of K starting from $K = 1$ until the value of s^2 shows a small decrease that fails to decline in substantial amount. This methods can easily be seen by plotting the value of variance estimated, s^2 versus the number of basis, K.

Figure 1.8 shows an **example** of s^2 against number of basis K plot, for $K = 1,\ldots,720$. Through a very sharp and close view that need to handle with care, $K = 121$ is chosen as the number of basis function since s^2 significantly fails to decline substantially after $K = 121$. There are more than one values show for lower s^2, but not selected as K.

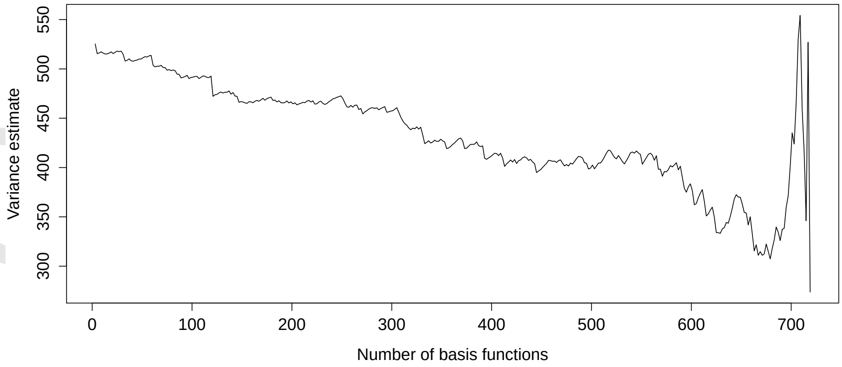

Figure 1.8: The relationship of number of basis K and the unbiased estimated of residual variance for $K = 1,\ldots,720$.

This is for the reason that, by adding more value of K possibly will not change the smoothness to a smoothest curve, yet will under smooth the data, Ramsay and Silverman (2005).

The relation between the number of basis function and the estimated variance besides, can be retrieved easily by using a deviance analysis, Jamaludin and Jemain (2011). The model in Equation (1.1.4) is approximate to a multiple of χ^2 with $T-K$ as a degree of freedom, Jamaludin and Jemain (2011). Thus, the analysis of deviance consists of p-value from χ^2 test of reduction in deviance for each K.

Table 1.1 shows an **example** of deviance analysis with several values of p-value of reduction in deviance for $K = 113, \dots, 129$. The analysis supposedly represents the p-value from $K = 1, \dots, 720$. The significant value is $K = 121$ as p-value is less than 0.05. This K which also has a largest reduction in deviance of 20.4477 and show that s^2 is fails to decline for the next value of K. Through these plotting and analysis of deviance approaches, an appropriate value of K can be obtained. It is important to select a suitable K in order to obtain a smooth function, since a very small value of K lead to over smooth the curve while a very large value of K gives an under smooth curve.

Table 1.1: Example of the analysis of deviance for respective value of K.

Basis,K	Estimated Variance,s^2	Reduction in Deviance	p-value
113	492.5549	0.3064	0.8579
115	491.3333	1.2216	0.5429
117	491.2973	0.0359	0.9821
119	492.6675	-1.3702	1.0000
121	**472.2198**	**20.4477**	$<$ **0.0001**
123	473.7527	-1.5329	1.0000
125	474.1958	-0.4430	1.0000
127	475.7003	-1.5046	1.0000
129	476.5167	-0.8164	1.0000

Smoothing penalties are introduced to reduce noise in measurements and very useful when the value of K is large. This study will be used both least squares and roughness penalty approaches in order to gain an appropriate results of smoothing. Roughness penalties is suit well for general problems in FDA and practical for an extensive range especially in smoothing problems. It also provides a good estimation in derivatives and can be represented by penalized residual sum of squares (PENSSE)

$$PENSSE = \sum_{t=1}^{T} (x_t - f_t)^2 + \lambda \int \left[f''(t) \right]^2 dt \qquad (1.1.5)$$

where $f^{''}(t)$ is a second derivative of function $f(t)$ which measures for a roughness of $f(t)$, and λ is a smoothing parameter that stabilized fit to the x_t and roughness.

The significant value of K define in the previous least squares approach with the respective value of basis function, ϕ_k is used as a reference value in the roughness penalty approach. In roughness penalty method, the parameter λ represents a smoothing parameter which identifies the amount of the smoothing that best represent K.

There are few conditions to be concerned in determining the value of λ. A very small value of λ lead the curve turn into more variable as the roughness consists of fewer penalty. For $\lambda \rightarrow 0$, then each of the observed value is equal to the fitted value, where $x_t = f_t$ for all t and the curve $f(t)$ is interpolate which will fit the data well as roughness is less. In contrast, if λ is very large, the curve is more smoother and at the same time less to fit the data. Whereas as $\lambda \rightarrow \infty$, there is no penalty in roughness that nearly results to the standard linear regression, Ramsay and Silverman (2005).

Figures 1.9 shows an **example** of the result for a small value of λ, for $t = 1, \ldots, 36$.

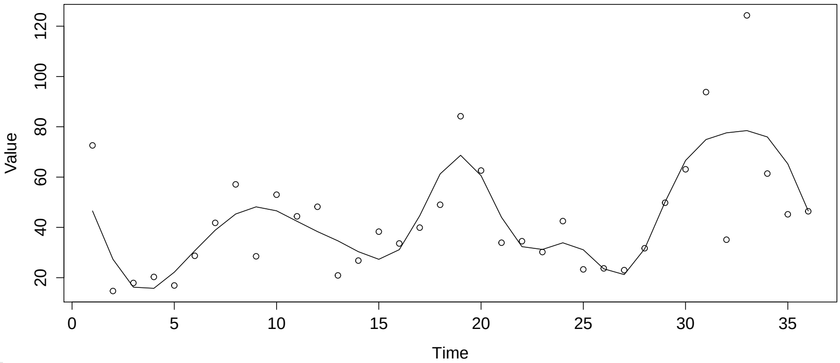

Figure 1.9: Example of the illustration of the curve for a small value of λ for $n = 36$.

The curve is slightly to have a smooth function, however it is more variable and more noises are taken into account since λ have too small value. Figure 1.10 depicts an **example** of the result when $\lambda \rightarrow 0$, for $t = 1, \ldots, 36$. The curve is not smooth as it approximately join the entire points in the data but then fit the data well.

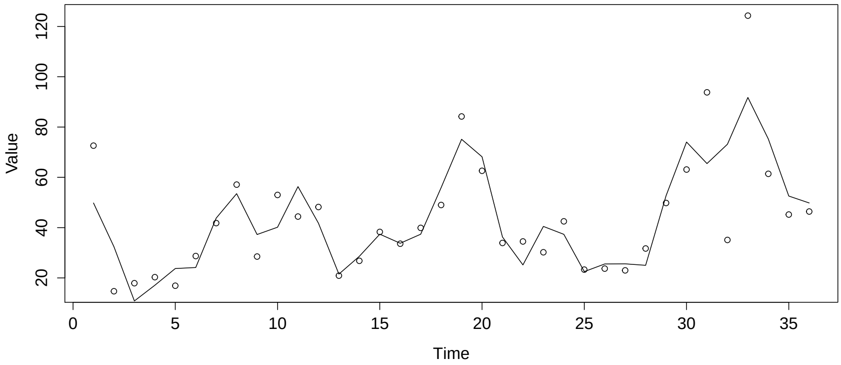

Figure 1.10: Example of the illustration of the curve for the value of $\lambda \to 0$ for $n = 36$.

Figure 1.11 shows an **example** of the result of a curve when λ has a large value, for $t = 1, \ldots, 36$. The function obtained seems illustrates a smooth curve, but more accurately the curve is slightly over smooth.

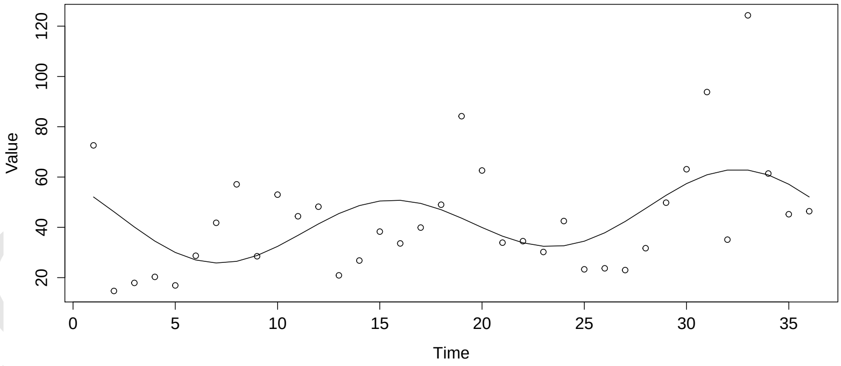

Figure 1.11: Example of the illustration of the curve for a large value of λ for $n = 36$.

Figure 1.12 displays an **example** of the result of a curve for $\lambda \to \infty$, for $t = 1, \ldots, 36$. As the λ value approach ∞, the function obtain a smooth curve, nevertheless the curve is approximate to a straight line. Therefore, it is important to chose the appropriate value of K and λ in oder to obtain for the best smooth curve which at the same time can fit the data well.

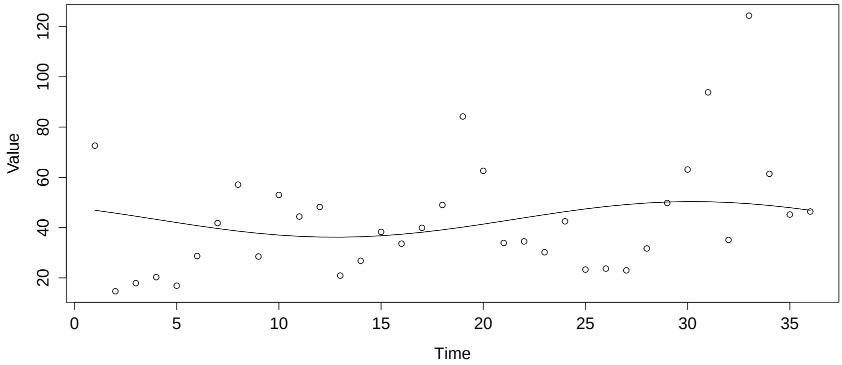

Figure 1.12: Example of the illustration of the curve for the value of $\lambda \to \infty$ for $n = 36$.

The generalize cross-validation (GCV) suggested by Craven and Wahba (1979), is used to determine the best value of the λ. The GCV can be calculated by the following formula

$$\mathrm{GCV}(\lambda) = \left(\frac{T}{T - df(\lambda)}\right)\left(\frac{SSE}{T - df(\lambda)}\right) \tag{1.1.6}$$

where $df(\lambda)$ is a degree of freedom for smooth function. Here, λ gives the smallest value of GCV will be chosen, Ramsay and Silverman (2005) and Ramsay et al. (2009). The smallest value of GCV that have to be chosen can be aided by plotting the $\log_{10}\lambda$ against $\log_{10}$ GCV, King (2014).

Some suitable values of λ are selected to setup precisely and will be tested in order to determine the value of GCV. As an example, Table 1.2 shows several values are setup for λ to obtain the values of GCV, together with the values of $\log_{10}\lambda$ and $\log_{10}$ GCV respectively. The smallest value of GCV is 521.2416 at $\lambda = 1$. The values of $\log_{10}\lambda$ and $\log_{10}$ GCV are use to plot the relationship between λ and GCV as an alternative to search for a smallest value of GCV, as displays in Figure 1.13.

There is condition where the least squares and roughness penalty give a slightly similar smooth function like in the Figure 1.14. This situation shows that fitting with least squares consider is an adequate result since there is no large lose in the fitted curve.

Table 1.2: Example of the analysis of deviance for respective value of K.

λ	GCV	$\log_{10}\lambda$	$\log_{10}$GCV
0.001	566.3580	-3	2.7531
0.01	557.1718	-2	2.7459
0.1	529.4524	-1	2.7238
1	**521.2416**	**0**	**2.7170**
10	523.9413	1	2.7192
100	532.3276	2	2.7261
1000	537.6095	3	2.7304
10000	538.3993	4	2.7311
100000	538.4820	5	2.7311

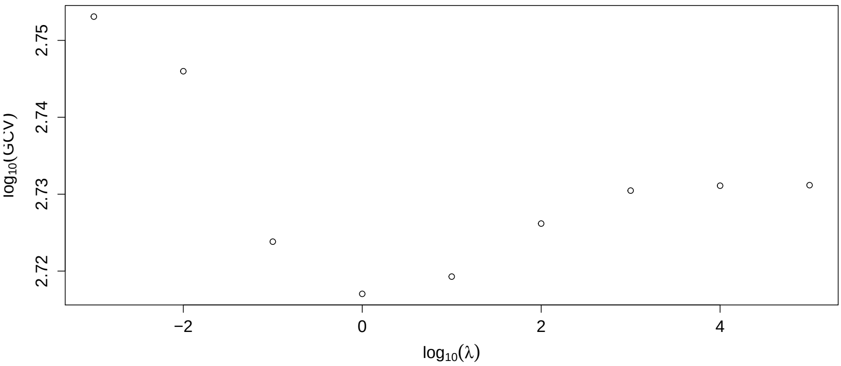

Figure 1.13: The relationship between $\log_{10}\lambda$ and $\log_{10}$ GCV. The value of $\log_{10}\lambda = 0$ result for the smallest value in GCV.

Nevertheless, some values of λ obtain from GCV leans toward an under smooth or over smooth the functions. In this case, the fundamental method of trial and error is needed to indicate the most appropriate λ through a self judgment, Yaraee (2011). Moreover, the value of K in the least squares approach are also difficult to be chosen. Thus, one believes that the best way in selecting the value of K and λ is to use the value which a fitted curve seems closer to the raw data.

A new selection model method in functional extreme data analysis is propose since the extreme data in FDA facing the same problem in selecting the value of K. This is due to a small data represents in extreme data. The proposed method is properly discussed in Chapter 5.

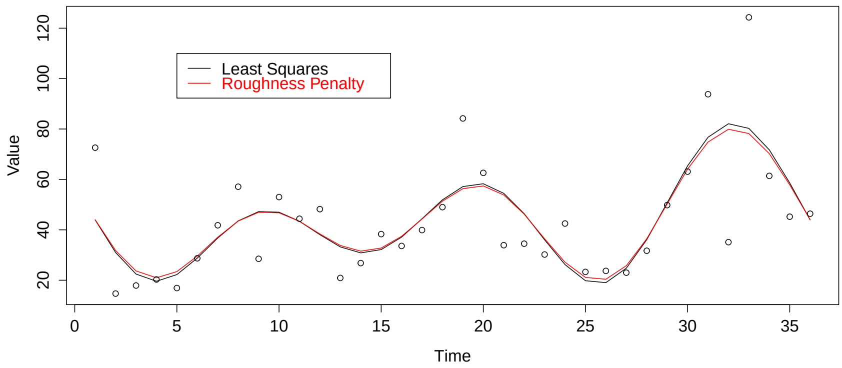

Figure 1.14: The smooth curve from least squares and roughness penalty method. The black line is a least squares smooth and the red line is a roughness penalty smooth for 36 observations with $K = 7$ and $\lambda = 0.1$.

1.1.2 Extreme Value Data

The extreme value data is fundamentally from extreme value theory with mathematical dominated by Leadbetter in 1980 and statistically explored by Gumbel in 1958. The details method and theory of extreme value can be found in Coles (2001) and Leadbetter and Rootzen (1988). This study will not discuss the detail of the theory and method in extreme value since the study simply implement the data from the extreme value approaches.

The tail of the distribution in classical statistics is an important thing to care of as it might influence the results of the analysis for most of the time, which can be categorized as short tail, long tail, heavy tail and etc. This tail can sometimes give a difficulty in the analysis that lead for other methods or approaches to be considered to solve for the problems occurred. On the other hand, the behavior of the tail distribution is an important thing to study. The extreme value data is a study that concerns for the tail distribution behavior either at the minimum or maximum values of $X_1, X_2, X_3, \ldots, X_n$ that are independent and identically distributed random variables, Haan and Ferreira (2006). Furthermore, extreme value data take into account for the large or small levels of process which involve the rare events, Coles (2001).

This study focus on maximum values as a data. The method of minimum values in the data will be not consider as a data. Let $X_1, X_2, X_3, \ldots, X_n$ a random variables of independent and identically distributed with distribution function F. The extreme value can be denoted as

$$M_n = \max\{X_1, \ldots, X_n\} \tag{1.1.7}$$

where X_i in time scale measurement for $i = 1, \ldots, n$ and M_n is the maximum value of observations over n time units, Coles (2001). As an example, if n is the number of observations in a months, then M_n is a monthly maximum value.

The extreme value data consists of three approaches in order to obtain the extreme data which can be listed as block maxima, r-largest order statistics and peak over threshold data as shown in Figure 1.15.

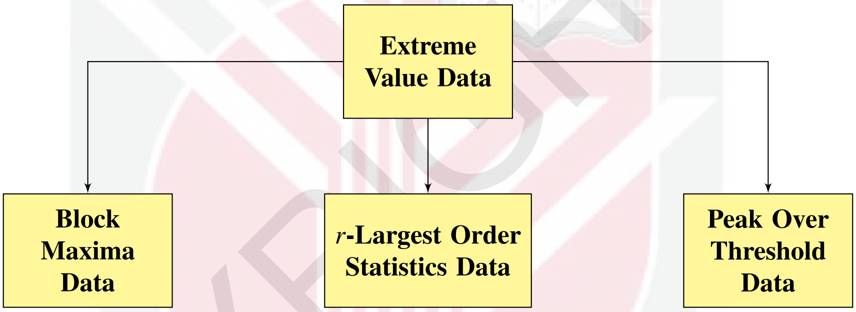

Figure 1.15: The diagram of extreme value data approaches which are block maxima, r-largest order statistics and peak over threshold data. (Coles, 2001)

The block maxima takes the data based on the highest value in every n number of observations, the r-largest order statistics choose the data based on the highest r order value in n observations, while peak over threshold data picks the data that fall above the threshold line.

These three extreme data approaches have the advantages and disadvantages according to the type of the data set used to be implemented on the analysis. Plus, not all the three approaches can be applied on the same data set where it is depends and can be influenced by the analysis to be carried out.

Block Maxima Data

The idea of extreme data is starts with the block maxima approach. This approach is obtained through the blocking time system of equal length of n, then will create m number of blocks which generates a block maxima series of $M_{n,1},\ldots,M_{n,m}$. The blocking system has a critical concern in choosing the size of the block. A small length of block lead to the extrapolation and estimation bias. While large length of block will give a few number of block maxima lead to large estimation variance. For that reason, Coles (2001) suggests to block the data by one year period of length, n or known as annual maxima.

Figures 1.16 and 1.17 show an **example** of the length of the block maxima which could affect the raise of the bias if too small or too large length of block will be chosen. As can be seen in Figure 1.16, the block has a small n where the extreme values from each block are too close that give an inaccurate result for estimation.

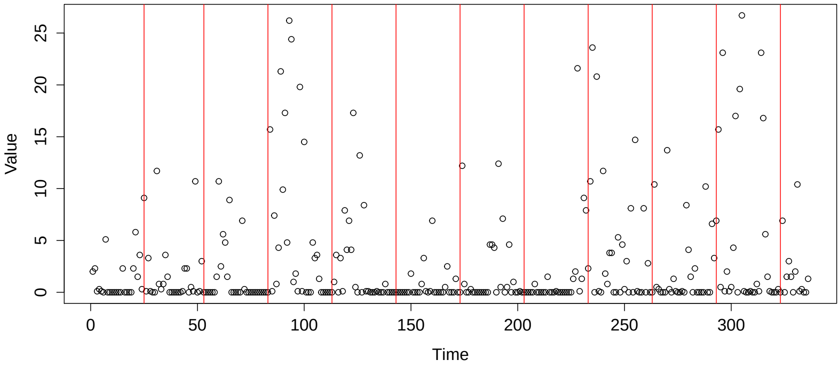

Figure 1.16: Example of too small n of blocking system of block maxima for 365 days.

Figure 1.17 shows an **example** of the block that has a large n which give a large distance of extreme value from every block that contribute to a large variation in estimation.

Thus, a big size of block gives a small number of data, while a too small size of block provides a lot of data which is close to each other. The notation for the value of block maxima, $M_{n,i}$ for $i = 1,\ldots,m$, then can be simplified into Z_i. The value taken from each block maxima again can be written as

$$M_{n,i} = \max\left\{X_{1,i}, \ldots, X_{n,i}\right\} \tag{1.1.8}$$
$$Z_i = M_{n,i}. \tag{1.1.9}$$

As the number of block represents as m, then for the maximum value from m block will be written as $Z_i, \ldots, Z_m$ for a correspond time length n.

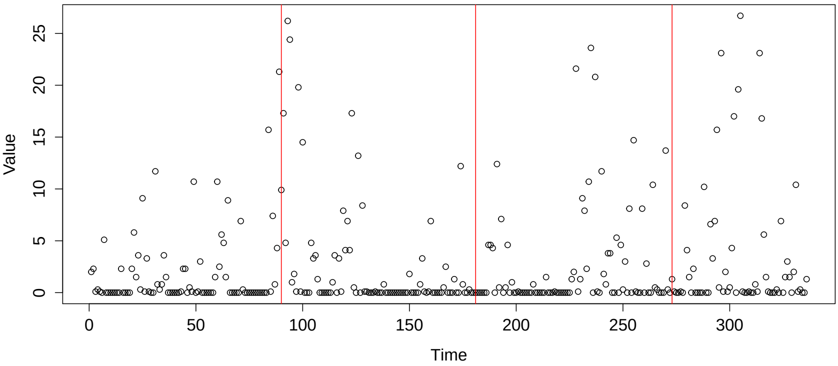

Figure 1.17: Example of too large *n* of blocking system of block maxima for 365 days.

r-largest Data

The block maxima limitation in obtaining a large number of data has come to a new search of extension for other approach. Taking more than one largest value from the block possibly will gain the amount of extreme data rather than taking only one maximum or minimum value from each block. Then, this approach is called as *r*-largest order statistics which is based on the behavior of the next largest values in the block. The other extreme order statistics of equal time length *n* in block *i* defined by

$$M_{n,i}^{(r)} = r\text{th largest of}\left\{X_{1,i}, \ldots, X_{n,i},\right\}$$

where k is the value of order statistics. Dissimilar with $M_{n,i}$ from block maxima, $M_{n,i}^{(r)}$ from r-largest order statistics consists of several values of order r. These values are influenced by each other as they connected by $M_{n,i}^{(1)} \geq M_{n,i}^{(2)} \geq M_{n,i}^{(3)} \ldots \geq M_{n,i}^{(r)}$.

The r-largest order statistics series are then can be obtained by blocking the data into m blocks. Determine the value of r to be taken from each block as r usually setting as equal for every block, unless there are less data in some blocks . Each block will be denoted as i and the values of largest r_i recorded as z_i which can be simplified as

$$M_{n,i}^{(r_i)} = z_i^{(1)}, \ldots, z_i^{(r_i)}, \qquad \text{for } \; i = 1, \ldots, m. \tag{1.1.10}$$

The appropriate choice of values of r is required to avoid other uncertainty to be occurred such as high variance for a small values of r and leading to bias for a large r. As usual practice, r is selected to be as large as possible, follow the model solving adequacy.

Figure 1.18 shows an **example** of how the 3^{rd} largest order statistics values are taken from each of the block. Then only the most three largest values from each block that represent by the solid black points will be treated as a data to be analyzed. This 3^{rd} largest data can be written as $M_{30,i}^{(3_i)} = \{z_i^{(1)}, z_i^{(2)}, z_i^{(3)}\}$ as $n = 30$.

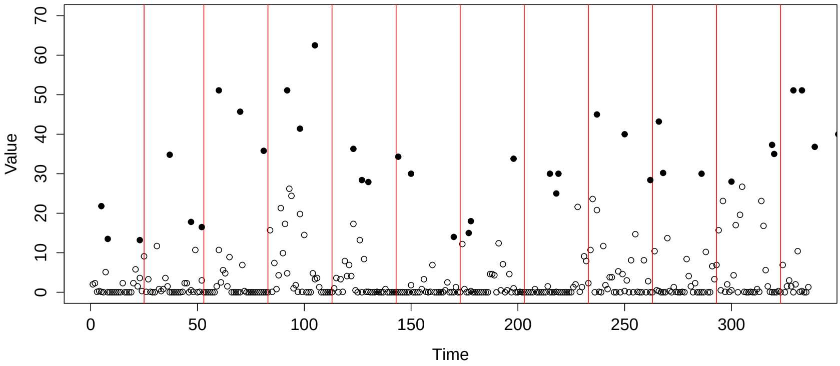

Figure 1.18: Example of 3^{rd} r-largest order statistics values taken from the each block. The solid points are selected as a data since the points represents the three highest values in each block.

The advantage of the r-largest approach is the size of the sample or data can be increased. The extreme data from block maxima and r-largest order statistics still have a limitation. The extreme data obtained are the highest values from each block. It is possibly that the other values from other block still represent a higher value than the selected values. It seems like a wasteful approach if other extreme data are available, Coles (2001). The third approach in obtaining the extreme data is applied and briefly introduce in the next subsection.

Threshold Data

Threshold approach is not follows the blocking procedure as practiced in block maxima and *r*-largest order statistics approaches. This contrast approach takes all the values above the threshold u, as an extreme data. The exceedances can be written as $\{x_i : x_i > u\}$ and denoted by $x_1, \ldots, x_k$. As well as block maxima approach, threshold approach has a problem in selecting the threshold value. This is due to the balance between the bias and the variance. A too low threshold will lead to the bias as the asymptotic basis of the model seems to be violate. While a few extreme data will be produce for a too high threshold which lead the variance to be high, Coles (2001).

These two problems clearly assist by graphical display as in Figures 1.19 and 1.20 for a well understanding. Figure 1.19 depict the **example** of too low threshold that involve too many exceedances treated as extreme data.

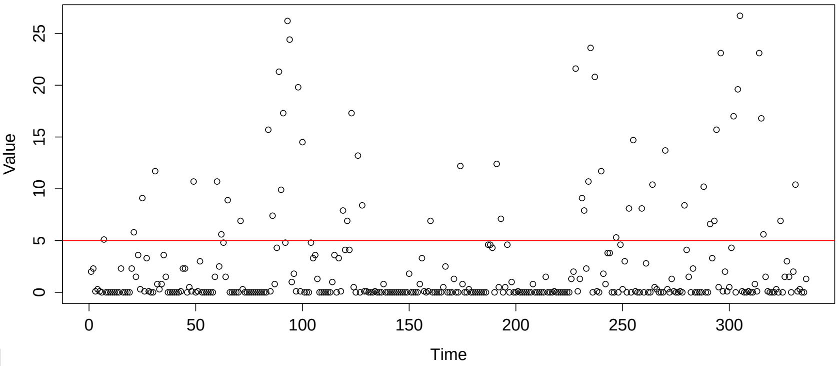

Figure 1.19: Example of extreme data from a too low threshold approach. The points over the threshold are treated as extreme data.

Figure 1.20 depict the **example** of too high threshold that involve too few exceedances treated as extreme data.

As a result, the threshold will be selected as low as possible where the model providing a reasonable approximation. There are few method used in selecting and determine an appropriate threshold such as by the graphical method or by following the some rules of thumb, Esther (2014). The details of these method will be explained in Chapter 4.

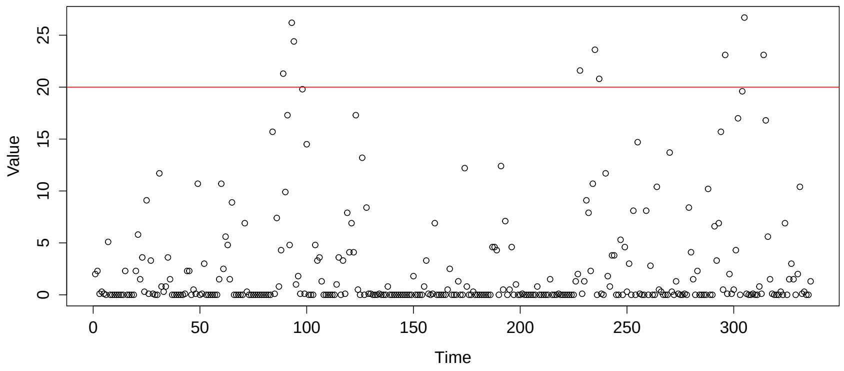

Figure 1.20: Example of extreme data from a too high threshold approach. The points over the threshold are treated as extreme data.

1.2 Motivation

Flood and drought are the calamity that can be cause by imbalance amount of rainfall and amount of run off in the certain area. Flood happen when the amount of rainfall is greater than the outflow of water, Jamaludin et al. (2011). Both disasters will give a great impact to agriculture sector and cause death if the disaster is seriously befallen. Hence, study the characteristics of the rainfall is one of the early preparation to overcome these disasters in order to reduce any lose, Jamaludin and Jemain (2011).

Markov chain models are often used to fit the rainfall occurrence. While two parameters gamma distribution, exponential distribution, weibull and lognormal are among the theoretical distribution used to fit the rainfall attribute, Ison et al. (1971), Cho et al. (2004), Bhakar et al. (2006).

Mixed-exponential is the best fit distribution for hourly rainfall data among exponential, gamma and weibull. While the mixed lognormal is the most appropriate distribution for describing the daily rainfall amount compare to lognormal and skew normal, Jamaludin and Jemain (2011).

On the other hand, time series analysis is often used to analyse environmental topics such as climate, hydrology, ozone level and etc. However, since the time series plots often resemble combined curves of acceleration and deceleration in either a positive or negative manner, one can conclude that some of the variation among curves could be explained at some level by derivatives, Ramsay and Silverman (2005).

The data in a functional form is more appropriate for derivatives rather than simply vectors of measurements over time, Ramsay and Silverman (2005). Besides, functional data analysis provides an extremely useful plot of velocity versus acceleration and also adds a modern twist on typical analysis, Allen (2011). Allen (2011) claims that functional data analysis compliments time series analysis fairly well. Nevertheless, functional data analysis offers a convenient technique in data arrangement and outcome where functional data analysis does not require for stationary assumption and not limited to the equally time space sample of data as applied in time series analysis, Ramsay and Silverman (2005).

1.3 Problem Statement

Functional data analysis has shown a various application in many fields by Ullah and Finch (2013), including in the environmental studies area. The early detection in Western region of air environmental studies are in the year 2007 by Meiring (2007) in Germany and Gao (2007) in Southern California, followed by Torres et al. (2011) in Spain while Shaadan et al. (2012) and Shaadan et al. (2014) from Malaysia represent the Eastern region.

In the recent years, the environmental application of functional data analysis has shown an increasing number of studies either in the Western or Eastern region. Jamaludin and Jemain (2011) starts to explore the rainfall data by using functional data analysis in Peninsular Malaysia, followed by Hamdan et al. (2013) and the latest by Wan and Jamaludin (2015). These three studies describe the characteristics of daily rainfall data by region and explore the similarity of the pattern for each region.

Many unexpected events occurred currently in the environment that need to be concerned as it cause a huge number of death and also damage to the affected areas especially by flooding events. Excessive rain is one of the floodplain which happen in almost of the country in the world including Malaysia at the end of the year 2014. The exploration on daily data is not appropriate to be applied to explore the behavior of these rare events.

King (2014) explores the functional data analysis on the climate change weather data instead of daily data. The climate change data also important in observing the changes in average weather condition but not involve the rare events such as Tsunami wave in Acheh 2004, a big flood in Malaysia in 2014, El Nino, La Nina phenomenon and etc.

Based on the literature review, all the previous studies on functional data analysis are concern only for daily rainfall data instead of extreme rainfall data that represent the rare events. Yet more studies of functional data analysis in environment especially for extreme events are needed to discover a huge and wider analysis which can not be done by discrete data.

In order to completely observe these rare events, a study of extreme data has to be done. The implementation of functional data analysis on extreme data can be a perfect match to analyse the rare events comprehensively. The aims of the study is to analyse the extreme data by using functional data analysis with the descriptive statistics exploration on rainfall data.

1.4 Objective

The study proposes a method development in functional extreme data analysis (FEDA) and the application on environmental data. The objectives of the study are to:

1. introduce functional data analysis (FDA) method to extreme data i.e. functional extreme data analysis (FEDA).

2. obtain smoother function in functional extreme data analysis (FEDA) by using Corrected Akaike Information Criterion (AICc) and Smoothing Index Rank (SIR).

3. develop tolerance bands for functional extreme data analysis (FEDA) based on bootstrapping method.

4. conduct an application of functional extreme data analysis (FEDA) on rainfall data.

1.5 Expected outcome

The uses of FDA in extreme data constitute the finest way in performing the rare events. The general expectation of this unification is to propose a new theoretical development for functional extreme data analysis (FEDA) for univariate cases. This method is extended to the threshold exceedances which the value for the mean is determined by the method of the rules of thumb. The results obtain will contribute for general features of FEDA that need an evaluation in preparing a fitted curve. The use of simulation data from generalized extreme value distribution and an application of a rainfall environmental data for both FEDA and threshold methods are also proposed to show and proved for the reliability of the outcome.

1.6 Significance of the study

The environmental data can be a very useful basis to study and monitor the specific critical area. This can be seen more clearly by the curve which represent the information in a different way and in a well view. Since many unexpected event over the world in this recent year that is caused by the unstable environment cyclic, these events recorded as rare events. These rare events might give a big impact to the world once it happened. Thus results from these types of data are greatly needed for a several countries and even to the world in estimating the future. The purpose of the thesis is to reduce damage from any natural disasters such as flooding, earthquake or huge wave by giving an alarm or alert on specific time. Study on extreme rainfall data allows decision makers to avoid or reduce flood which can affect on loses of economy, to construct water management system and also to advice insurance against water damage, Ender and Tong (2014).

1.7 Scope of the study

There scope of the study and its relevance are as follows:

1. The study explores the functional data analysis on extreme data in descriptive statistics stage as the extreme data are different with the normal data. Besides, there are no references for functional data analysis on extreme data except one reference on climate change data.
2. The study uses the fourier basis function system for functional data analysis on extreme data since it is periodic data.
3. The study uses to block the raw daily data by month for each year in application to extreme rainfall data since the study have the whole complete daily data set as suggested by Coles (2001).
4. The study applies the functional data analysis method on extreme rainfall data to five stations in Peninsular Malaysia to represent the north, south, east, west and middle areas for 30 years period of data.
5. The study uses only one parameter which is rainfall in application to real data.

1.8 Overview

This thesis chapters are divided into seven chapters start from introduction of the study to the recommendation for future work with each chapter have their own importance to the thesis. Chapter 1 introduces briefly of the background of the study related to the theory of FDA and method in determine the extreme data. Specifically, the objectives of the study, how this thesis is motivated, what are the significance of this study that will contribute more or less to the country otherwise to the world, the expected outcome of the study and what are the limitation this study facing are stated here.

Chapter 2 represents some review on functional data analysis in environmental application with a table of summary comparing previous studies in functional data analysis application on rainfall data also provided. Chapter 3 discusses an implementation of methodology between FDA and extreme data from block maxima and *r*-largest approaches. As well as simulation and real data are tested along with an appropriate results and conclusion. An implementation of methodology between FDA and threshold data are extended in Chapter 4 with the results and conclusion of the simulation and application to real data presented in this chapter.

Chapter 5 defines a new approach for model selection in FDA for extreme data by introducing the corrected Akaike Information Criteria (AICc) and Smoothing Index Rank (SIR). A new concept in finding a tolerance bands for extreme data in FDA is discussed in Chapter 6. The additional of the latest set of rainfall data from five stations are performed in Chapter 7 and the recommendation for the future work is presented in Chapter 8.

REFERENCES

Ainsworth, L., Routledge, R., and Cao, J. (2011). Functional Data Analysis in Ecosystem Research: The decline of Oweekeno Lake Sockeye Salmon and Wannock River Flow. *Journal of Agriculture, Biological and Environmental Statistics*, 16:282–300.

Akaike, H. (1992). Information Theory and an Extension of the Maximum Likelihood Principle. *Springer*, 1:610–624.

Allen, J. (2011). *Comparison of Time Series and Functional Data Analysis for the Study of Seasonality*. Phd dissertation, Department of Mathematics, East Tennessee State University.

Andreev, V. O., Tinykov, S. E., Ovchinnikova, O. P., and Parahin, G. P. (2012). Extreme Value Theory and Peaks Over Threshold Model in the Russian Stock Market. *Journal of Siberian Federal University*, 1:111–121.

Babura, B. I. (2017). *Modified Boxplot and Stairboxplot for Generalized Extreme Value Distribution*. Phd dissertation, Institute for Mathematical Research, Universiti Putra Malaysia.

Bertoldo, S., Lucianaz, C., and Allegretti, M. (2015). Extreme Rainfall Event Analysis Using Rain Gauges in a Variety of Geographical Situations. *Atmospheric and Climate Sciences*, 5:82–90.

Bhakar, S. R., Bansal, A. K., Chhajed, N., and Purohit, R. C. (2006). Frequency Analysis of Consecutive Days Maximum Rainfall at Banswara, Rajasthan, India. *Journal of Engineering and Applied Science*, 1:64–67.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, second edition.

Cheng, Y., Chiu, Y., Wang, H., Chang, F., Chung, K., Chang, C., and Cheng, K. (2013). Using Akaike information criterion and minimum mean square error mode in compensating for ultrasonographic errors for estimation of fetal weight by new operators. *Taiwanese Journal of Obstetrics & Gynecology*, 52:46–52.

Chernick, M. R. and LaBudde, R. A. (2011). *An Introduction to Bootstrap Methods with Applications to R*. Wiley.

Cho, H. K., Bowman, K. P., and North, G. R. (2004). A Comparison of Gamma and Lognormal Distribution for Characterizing Satellite Rain Rates from the Tropical Rainfall Measuring Mission. *Journal of Applied Meteorology*, 43:1586–1597.

Clark, M. and Anfinson, C. (2011). *Intermediate Algebra:Connecting concept through application*. Charlie Van Wagner, second edition.

Coles, S. (2001). *An introduction to Statistical Modeling of Extreme Values*. Springer, London.

Craven, P. and Wahba, G. (1979). Smoothing Noisy Data with Spline Functions:Estimating the Correct Degree of Smoothing by the Methods of Generalized Cross Validation. *Numerische Mathematik Springer*, 31:377–403.

Delson, C. and Retius, C. (2015). Modeling of extreme minimum rainfall using generalized extreme value distribution for Zimbabwe. *South African Journal of Science*, 111(9/10):8 pages.

Douglas, B. C., Fraley, C., Charles, C. G., and Ramsay, J. (2005). *S+ Functional Data Analysis*. Springer, New York.

Efron, B. (1979). Bootstrap methods: another look at the jacknife. *Annal Statistics*, 7:1–26.

Eischeid, J. K., Baker, C. B., Karl, T. R., and Diaz, H. F. (1995). The quality control of long term climatological data using objective data analysis. *Journal of Applied Meteorology*, 34:2787–2795.

Ender, M. and Tong, M. (2014). Extreme Value Modeling of Precipitation in Case Studies for China. *International Journal of Scientific and Innovative Mathematical Research*, 2:23–36.

Esther, B. (2014). Peak -Over-Threshold Modelling of Environmental Data. Internet. Retrieved Nov 11th, 2016.

Gao, H. (2007). Day of week effects on diurnal ozone/NOx cycles and transportation emissions in Southern California. *Transportation Research Part D*, 12:292–305.

Goldsmith, J., Greven, S., and Crainiceanu, C. (2013). Corrected confidence bands for functional data using principal components. *Biometrics*, 69(1):41–51.

Haan, L. and Ferreira, A. (2006). *Extreme Value Theory:An Introduction*. Springer, New York.

Hamdan, M., Jamaludin, S., and Jemain, A. (2013). Functional Data Analysis Technique on Daily Rainfall Data : A Case Study at North Region of Peninsular Malaysia. *MATEMATIKA*, 29:233–240.

Hassan, H., Salam, N., and Adam, M. B. (2013). Modelling Extreme Temperature in Malaysia Using Generalized Extreme Value Distribution. *International Journal of Mathematical, Computational, Physical, Electrical and Computer Engineering*, 7(6):983–989.

Hershfield, D. (1983). A Comparison of three Generalized Extreme-Value Rainfall Studies. *Archives For Meteorology, Geophysics, and Bioclimatology*, 33:251–260.

Hurvich, C. M. and Tsai, C. H. (1989). Regression and time series model selection in small samples. *Biometrika*, 76:297–307.

Husain, Q. (2017). *Modifications of Tukey's Smoothing Techniques for Extreme Data*. Phd dissertation, Faculty of Science, Universiti Putra Malaysia.

Ison, N. T., Feyerherm, A. M., and Bark, L. D. (1971). Wet Period Precipitation and the Gamma Distribution. *Journal of Applied Meteorology*, 10:658–665.

Jamaludin, S., Deni, S. M., and Jemain, A. (2008). Revised Spatial Weighting Methods for Estimation of Missing Rainfall Data. *Asia Pacific Journal of Atmospheric Sciences*, 44:93–104.

Jamaludin, S. and Jemain, A. (2011). Comparing rainfall patterns between regions in Peninsular Malaysia. *Journal of Hydrology*, 411:197–206.

Jamaludin, S., Kong, C., Yusof, F., and Foo, H. (2011). Introducing the Mixed Distribution in Fitting Rainfall Data. *Open Journal of Modern Hydrology*, 1:11–22.

Keser, I., Kocakok, I., and Sehirlioglu, A. (2016). A New Descriptive Statistics for Functional Data: Functional Coefficient of Variation. *Alphanumeric Journal*, 4(2):1–10.

King, K. (2014). *Functional Data Analysis With Application to United States Weather Data*. Honors college theses, Mathematics and Statistics, UVM College.

Klassen, K. J. (1997). *Simultaneous Management of Demand and Supply in Services*. Phd dissertation, Faculty of Management, The University of Calgary.

Krishnamoorthy, K. and Mathew, T. (2009). *Statistical tolerance regions: theory, applications, and computation*. John Wiley & Sons, New Jersey.

Lan, F., Micheal, M., and Szu, H. (2013). Statistical Modeling of Recent Changes in Extreme Rainfall in Taiwan. *International Journal of Environmental Science and Development*, 4(1):52–55.

Leadbetter, M. and Rootzen, H. (1988). Extremal Theory for Stochactics Processes. *The Annals of Probability*, 2:431–478.

Meiring, W. (2007). Oscillations and Time Trends in Stratospheric Ozone Levels : A Functional Data Analysis Approach. *Journal of the American Statistical Association*, 102:788–802.

Michael, R. C. (2008). *Bootstrap Methods:A Guide for Practitioners and Researchers*. Wiley.

Michael Snipes, D. and Taylor, C. (2014). Model Selection and Akaike Information Criteria: An Example from Wine Ratings and Prices. *Wine Economics and Policy*, 3:3–9.

Okorie, I. and Akpanta, A. (2015). Threshold Excess Analysis of Ikeja Monthly Rainfall in Nigeria. *International Journal of Statistics and Applications*, 5:15–20.

Paulhus, J. L. H. and Kohler, M. A. (1952). Interpolation of missing precipitation records. *Monthly Weather Review*, 8:129–133.

Penny, W. (2012). Comparison Dynamic Causal Models using AIC, BIC and Free Energy. *Neuroimage*, 59:319–330.

Posada, D. and Buckley, T. (2004). Model Selection and Model Averaging in Phylogenetics: Advantage of Akaike Information Criterion and Bayesian Approaches Over likelihood Ratio Tests. *Society of Systematic Biologists*, 53:793–808.

Ramsay, J., Hooker, G., and Graves, S. (2009). *Functional Data Analysis with R and MATLAB*. Springer, New York.

Ramsay, J. and Silverman, B. (2002). *Applied Functional Data Analysis:Methods and Case Studies*. Springer, New York.

Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis*. Springer.

Rathnayake, L. N. and Choudhary, P. K. (2016). Tolerance bands for functional data. *Biometrics*, 72(2):503–512.

Roberto, V., Georg, G., and Manfred, S. (2005). Functional Principal Component Analysis of fMRI Data. *Human Brain Mapping*, 24:109–129.

Ryde, J. (2005). Extreme-Value modelling: A preliminary analysis of monthly precipitation at Havana. *Trabajos Teoricoexperimentales*, 27:20–23.

Saralees, N. and Dongseok, C. (2007). Maximum daily rainfall in South Korea. *Journal of Earth System Science*, 116:311–320.

Shaadan, N., Deni, S. M., and Jemain, A. A. (2012). Assessing and comparing PM_{10} Pollutant Behaviour using Functional Data Approach. *Sains Malaysiana*, 41:1335–1344.

Shaadan, N., Jemain, A. A., and Deni, S. M. (2014). Data Preparation for Functional data Analysis of PM10 in Peninsular Malaysia. AIP Publishing.

Simanton, J. R. and Osborn, H. B. (1980). Reciprocal distance estimate of point rainfall. *Journal of the Hydraulics Division*, 106:1242–1246.

Snipe, M. and Taylor, D. (2014). Model selection and Akaike Information Criteria: An example from wine ratings and prices. *Wine Economics and Policy*, 3:3–9.

Song, F., Sarales, N., and Qi, H. (2007). Modeling Annual Extreme Precipitation in China Using the Generalized Extreme Value Distribution. *Journal of the Meteorological Society of Japan*, 85:599–613.

Srivastava, M. and Kubokawa, T. (2008). Akaike Information Criterion for Selecting Components of the Mean Vector in High Dimentional Data with Fewer Observations. *J. Japan Statist. Soc*, 38:259–283.

Tabios, G. Q. and Salas, J. D. (1985). A Comparative Analysis of Techniques for Spatial Interpolation of Precipitation. *Water Resources Bulletin*, 21:365–380.

Temiyasathi, C. (2008). *Functional Data Analysis for Environmental and Biomedical Problems*. Phd dissertation, Faculty of the Graduate School, The University of Texas.

Torres, J. M., Nieto, P. G., Alejano, L., and Reyes, A. (2011). Detection of outliers in gas emission from urban areas using functional data analysis. *Journal of Hazardous Materials*, 186:144–149.

Tronci, N., Molteni, F., and Bozzini, M. (1986). A comparison of local approximation methods for the analysis of meteorological data. *Archives for Meteorology, Geogphysics, and Bioclimatology*, 36:189–211.

Ullah, S. and Finch, F. (2013). Applications of functional data analysis: A systematic review. *BMC Medical Research Methodology*, 13:1471–2288.

Wan, W. and Jamaludin, S. (2015). Smoothing Wind and Rainfall Data through Functional Data Analysis Technique. *Jurnal Teknologi*, 74:105–112.

Wang, J. L., Chiou, J. M., and Muller, H. G. (2015). Review of Functional Data Analysis. *Annual Reviews Statistics*, 1:1–41.

Xia, Y., Fabian, P., Stohl, A., and Winterhalter, M. (1999). Forest Climatology: estimation of missing values for Bavaria, Germany. *Agriculture and Forest Meteorological*, 96:131–144.

Yaraee, K. (2011). *Functional Data Analysis with Application to MS and Cervical Vertebrae Data*. Masters thesis, Department of Aerospace Engineering, Faculty of Engineering, University of Alberta.

Young, K. C. (1992). A three way model for interpolating monthly precipitation values. *Monthly Weather Review*, 120:2561–2569.