

Smoothing Wind and Rainfall Data through Functional Data Analysis Technique

W. I. Wan Norliyana, Jamaludin Suhaila*

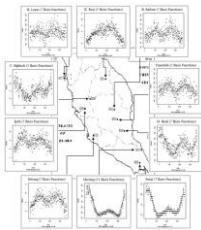
Department of Mathematical Sciences, Faculty of Science, Institute of Environmental & Water Resource Management, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia

*Corresponding author: suhailasj@utm.my

Article history

Received :21 May 2014
Received in revised form :
18 January 2015
Accepted :15 March 2015

Graphical abstract



Abstract

The pattern of wind and rainfall throughout Peninsular Malaysia are varied from one region to another, because of strong influences from the monsoons. In order to capture the wind and rainfall variations, a functional data analysis is introduced. The purpose of this study is to convert the wind and rainfall data into a smooth curve by using functional data analysis method. Fourier basis is used in this study since the wind and rainfall data indicated periodic pattern. In order to avoid such overfitting data, roughness penalty is added to the least square when constructing functional data object from the observed data. Result indicated that if we use a small number of bases functions, the difference is very small between with and without roughness penalty, showing that it is safer to smooth only when required. However, when a large basis function is employed, the roughness penalty should be added in order to obtain optimal fit data. Based on the contour plot of correlation and cross-correlation functions of wind and rainfall data, the relationship between both climate functions could be determined.

Keywords: Basis function; smoothing curve; roughness penalty; functional data; fourier basis

Abstrak

Corak angin dan hujan di Semenanjung Malaysia berubah mengikut wilayah disebabkan pengaruh kuat dari monsun. Bagi menerangkan variasi angin dan hujan, analisis data fungsi diperkenalkan. Tujuan kajian ini adalah untuk menukarkan data angin dan hujan ke dalam suatu fungsi licin dengan menggunakan kaedah analisis data fungsi. Basis Fourier digunakan dalam kajian ini kerana data angin dan hujan menunjukkan corak bermusim. Untuk mengelak data terlebih padanan, penalti kasar dikenakan ke atas kaedah kuasa dua dalam membentuk data fungsi dari data cerapan. Keputusan menunjukkan jika menggunakan bilangan fungsi basis yang kecil, perbezaan yang sangat sedikit dapat dilihat di antara kesan penalti kasar dengan tiada penalti kasar, maka lebih selamat untuk melicinkan sesuatu fungsi bila perlu. Walau bagaimanapun, jika basis yang besar digunakan, penalti kasar perlu dikenakan untuk menghasilkan padanan yang optimal. Berdasarkan, plot kontur korelasi dan korelasi silang fungsi angin dan hujan, hubungan antara kedua-dua fungsi iklim tersebut dapat ditentukan.

Kata kunci: Fungsi basis; lengkung licin; penalti kasar; data fungsi; basis Fourier

© 2015 Penerbit UTM Press. All rights reserved.

1.0 INTRODUCTION

Functional Data Analysis (FDA) develops fast in statistics area with the aim of estimating a set of related functions or curves rather than focusing on a single entity. The information about FDA such as the slopes, curvatures, and other characteristics are available based on the intrinsically smooth curves built up through FDA. The basic idea of FDA is to express discrete observations arising from time series into a functional data that represents the entire measured function as a single observation, and to draw modeling and make inference based on the collection of a functional data by applying statistical concept from univariate or multivariate data analysis.

Recently, FDA has been gaining momentum in many fields such as in medicine, biomedicine, public health, biological sciences, biomechanics, environmental science, and economics. For example in biomedical, a functional analysis of variance which is part of functional linear model has been employed to compare several groups in the experimental cardiology (e.g. Ferraty *et al.*, 2006; Cuevas *et al.*, 2003) meanwhile Ratcliffe *et al.* (2002) applied singular longitudinal analysis with functional regression to periodically stimulated foetal heart rates. Based on their results, the functional model for the stimulated foetal heart rates represents the enhancement for the best standard linear regression model. On the other hand, Nikitovic (2011) and Hyndman and Booth (2008) use

FDA to forecast demographic rates and the structure of the population while in the area of meteorology, Suhaila and Jemain (2009) introduced FDA to convert rainfall observations into a smoothing rainfall curve which was then used in comparing the climate rainfall patterns between regions.

Generally, the wind in Malaysia is light and varies; however, some periodic changes in the wind flow patterns could influence the rainfall distribution. A strong wind is expected to bring heavy rainfall at the location. Our main objective in this study is to use FDA technique in representing the rainfall and wind data in the form of smoothing curves since wind and rainfall data are recorded as daily observations at a discrete time interval. FDA is used to represent the data in a way that could give information on the pattern and variation of the data and make use of the information in the slopes and curvatures of curves that are reflected in their derivatives. The smooth curve from FDA can then be used to compare rainfall and wind variability between regions. Contour plot of bivariate rainfall and wind could establish the relationship between both smoothing climate variables.

Therefore, a functional data analysis will be carried out to examine the changes and variability for both climate variables and establish the functional relationship between them. The outcome of this study is expected to be useful to policy makers, climatologist, and water resource planners dealing with climate change for the sustainable development and planning of water resources.

2.0 STUDY AREA AND DATA

Peninsular Malaysia is located between 1° 7' North latitude and 100° 103' East longitude. Peninsular Malaysia has several types of landscapes of its certain latitude and longitude measurements which has tropical weather and is affected by monsoonal climate.

The wind speed at the east coast of Peninsular is mostly influenced by the northeast monsoon (NEM) which occurs from November to February, while the southwest monsoon (SWM) may possibly influenced the wind speed at the west coast of Peninsular between May and August. Based on the daily rainfall distribution, Suhaila *et al.* (2011) used four rainfall regions in Peninsular Malaysia which are the Northwest, West, Southwest and East. The east coast of Peninsular Malaysia experiences heavy rainfalls during the NEM. Heavy rains are also expected during the two inter-monsoons: March to April and September to October. On the other hand, the stations or areas which are sheltered by mountain ranges are relatively free from those monsoons. It is best to distinguish the rainfall distribution of the stations according to seasons.

Meteorological data were obtained from Malaysian Meteorological Services (MMS). The wind data is the speed of wind in meters per second meanwhile rainfall data is in millimeters per day for a period of 25 years from 1985 to 2009. The list of stations is provided in Table 1.

Table 1 The list of ten stations with their geographical coordinates

Code	Stations	Latitude	Longitude
S01	Kuala Krai	5.45	102.30
S02	Batu Embun	4.15	102.75
S03	Temerloh	3.70	102.94
S04	Muadzam Shah	3.35	103.25
S05	Mersing	2.45	103.83
S06	Senai	1.63	103.67
S07	Bayan Lepas	5.30	100.27
S08	C. Highlands	4.60	101.50
S09	Ipoh	4.57	101.10
S10	Subang	3.12	101.55

3.0 METHODOLOGY

This section is divided into two main subsections. Method of finding suitable basis functions will be discussed in the first subsection. The method of least squares is used to estimate the parameters of the basis function. A functional descriptive statistics based on mean, standard deviation and correlation of functional data will be described in the second sub-section. Contour plots are used to describe the relationship between climates variables.

3.1 Basis Function

There are two methods that are normally used in representing the functional data, namely, smoothing and interpolation. If the discrete values are assumed to be errorless, the process involves interpolation method. But if they are some observational errors that need to be removed, the transformation from discrete data to function may require smoothing. The first step in FDA is to create a set of bases functions which used to convert the discrete values into a smooth curve. A set of linear combinations of bases functions is used in representing functions, which is given as

$$x(t) = \sum_{k=1}^K c_k \phi_k(t) \tag{1}$$

where c_k refers to the basis coefficient, ϕ_k is the known basis function while K is the size of the maximum basis required.

Fourier series give the best known basis function for periodic data, which can then be written as;

$$x(t) = c_0 + c_1 \sin \omega t + c_2 \cos \omega t + \dots \tag{2}$$

defined by the basis function $\phi_0(t) = 1$, $\phi_{2k-1}(t) = \sin k\omega t$, $X_{2k}(t) = \cos k\omega t$ with $t = t_1, \dots, t_T$. The basis is periodic, and the constant ω is related to the period T by the relation $\omega = \frac{2\pi}{T}$.

After the first constant basis function, Fourier basis functions are arranged in successive sine and cosine pairs.

Let y_i be the observed discrete data points for $i = 1, 2, \dots, T$, collected over a continuum time t . Assuming that there is a reasonably smooth function $x(t)$ that gives rise to those discrete points. Defining the model as

$$y_i = x(t_i) + \varepsilon_i. \tag{3}$$

with mean zero and constant variance σ^2 , the errors, ε_i are assumed to be independent and normally distributed.

In order to estimate a reasonably smooth function $x(t)$, the most well-known method of Least Squares estimation (SSE) is used. When the function $x(t)$ is defined in terms of the basis function expansion in Equation (1), the coefficients of the expansion, c_k are determined through the least squares method by minimizing the sum of squared residuals:

$$SSE = \sum_{i=1}^T (y_i - x(t_i))^2 \tag{4}$$

with y_i represent the original observed data and $x(t_i)$ is the fitted smooth data.

Larger values of K basis functions will tend to undersmooth or overfit the data (Ramsay *et al.*, 2009). Therefore, when a large number of basis functions are used, a more powerful method of smoothing called roughness penalty is introduced. The basic idea of the roughness penalty approach is similar to the least square with an additional of a penalty term in Equation (4) and multiplied by a smoothing parameter which plays the role of penalizing the roughness to produce a better result. The penalized sum of squares (PENSSE) is defined as

$$PENSSE = \sum_{i=1}^T (y_i - x(t_i))^2 + \lambda \int (x''(t))^2 dt. \quad (5)$$

The smoothing parameter λ controls a compromise between the fit to the data and the variability in the function. Large values of λ will increase the amount of smoothing. However, to determine the best value for smoothing parameter λ , generalized cross-validation (GCV) is applied and is defined as

$$GCV(\lambda) = \left(\frac{T}{T - df(\lambda)} \right) \left(\frac{SSE}{T - df(\lambda)} \right) \quad (6)$$

where $df(\lambda)$ refers to the number of degrees of freedom and T is the number of observations. Several values of λ are tested, and value of λ which gives the smallest GCV is used.

3.2 Descriptions of Functional Data

The descriptive statistics such as mean, variance, standard deviation, covariance and correlation are estimated for functional data. The correlation function, describe the relationship between times for a climate variable while the cross-correlation function is used to determine the dependency between climate variables.

Let $x_i, i = 1, \dots, N$, be a sample of curves or functions fits to data. The mean and variance are given as below; with the standard deviation function is the square root of the variance function;

$$\bar{x} = N^{-1} \sum_{i=1}^N x_i(t) \quad (7)$$

$$var_X(t) = (N - 1)^{-1} \sum_{i=1}^N [x_i(t) - \bar{x}(t)]^2. \quad (8)$$

The covariance function summarizes the dependence of records across difference argument values, and is computed for all t_1 and t_2 by

$$cov_X(t_1, t_2) = (N - 1)^{-1} \sum_{i=1}^N \{x_i(t_1) - \bar{x}(t_1)\} \{x_i(t_2) - \bar{x}(t_2)\}. \quad (9)$$

Meanwhile, the associated correlation function is given as

$$corr_X(t_1, t_2) = \frac{cov_X(t_1, t_2)}{\sqrt{var_X(t_1)var_X(t_2)}}. \quad (10)$$

Then, the dependency between variable can be quantified by the cross-covariance function

$$cov_{X,Y}(t_1, t_2) = (N - 1)^{-1} \sum_{i=1}^N \{x_i(t_1) - \bar{x}(t_1)\} \{y_i(t_2) - \bar{y}(t_2)\} \quad (11)$$

Using Equation (11), the cross-correlation function is given as

$$corr_{X,Y}(t_1, t_2) = \frac{cov_{X,Y}(t_1, t_2)}{\sqrt{var_X(t_1)var_Y(t_2)}}. \quad (12)$$

Contour plots display the contour line for a function of two climate variables between times and variables.

4.0 RESULT AND DISCUSSION

This section is divided into three main sub-sections. In the first sub-section, the number of bases functions required for each region will be identified and the results will be validated by examining the residuals to obtain an optimal fit to data. Smoothing the functional data with and without roughness penalty will be investigated in the second sub-section. The third sub-section will summarize the

pattern of wind-rainfall data using the functional descriptive statistics and establishing the relationship between climate variables based on correlation values.

4.1 Identifying the Number of Basis Functions

In this study, FDA technique involves building a functional data object from the observations that provide information on the pattern of the data. By comparing the deviance between the estimated and the observed values, we can determine the number of bases functions that best described the wind and rainfall data.

Table 2 illustrates the example showing the analysis of deviance of mean wind speed for Batu Embun and mean rainfall for Muadzam Shah. Low p-value indicates the possibility of rejecting the null hypothesis. Based on those results, there is evidence to show that seven basis functions with three harmonics are required in describing mean daily wind speed for Batu Embun while nine basis functions which represent four harmonics are sufficient in describing mean daily rainfall for Muadzam Shah. Figure 1 (wind speed) and Figure 2 (rainfall) display the resulting smooth curves based on the number of basis functions that have been obtained for ten stations. The wind and rainfall patterns are compared regarding to the smoothing curves of each station.

As shown in Figure 1, five and seven bases are required to describe the variation in wind speed for most of the stations. The highest number of bases functions is observed for Mersing in which eleven bases are required to describe the variation of wind speed of the station. High mean speed is observed during the northeast monsoon months (Nov to Feb) for stations Muadzam Shah, Mersing and Senai while sudden drops in the speed values are observed between April and October. Conversely, for other stations such as Kuala Krai, Batu Embun, Temerloh, Ipoh and Subang, high mean speed is recorded in April and October during the inter-monsoon months and slightly dropped during November to February. Different patterns are achieved for each station may possibly due to the factors such as geographical locations, distance from the sea and monsoons influence.

Figure 2 displays the smoothing rainfall curves with their number of bases functions of each station. There are different rainfall patterns are found between stations at east Peninsular and west Peninsular. As shown in Figure 2, a unimodal rainfall pattern is displayed for Mersing station at the east Peninsular. The highest seasonal rainfall peak is observed during the Northeast monsoon months while low rainfall values are recorded by the rest of the months. On the other hand, bimodal rainfall patterns are displayed by most of the stations at west Peninsular. The first seasonal rainfall peaks are observed in April to May while the second peak is in Sept to Oct. The second inter-monsoon is found to be wetter than in the first inter-monsoon. In comparison, the period from mid January to early February and from mid June to mid August are considered as dry periods for stations in the Western region.

4.2 Smoothing With and Without Roughness Penalty

For the smoothing data, we can sometimes get good results, by keeping the small number of bases functions related to the amount of data being approximated. On the other hand, to obtain an optimal fit to data, the other strategy aims by employing a powerful basis expansion. Roughness penalty may be added to the least square when constructing functional data object from the observed data. This approach allows finer control over the amount of smoothing. Based on the minimum GCV, parameter λ is chosen.

In order to compare the smoothing with or without roughness penalty, Temerloh and Cameron Highlands stations are taken as examples. Several values of λ are tested and the GCV values are obtained as shown in Table 3. Based on the minimum values of

GCV, the value of lambda is shown as $\lambda = 1e7$. The degrees of freedom for wind and rainfall data, are given as $df(\lambda) = 4.98 \approx 5$, which is equivalent to the number of bases functions that are used. Figure 3 shows the smooth functional data for Temerloh and Cameron Highlands with and without roughness penalty by using

small basis functions. As we can see from both stations, the differences of smoothing functions are too small in which the graphs seem to be overlapping with each other. This shows that by smoothing the functional data without a roughness penalty, we will not lose much and it is safer to smooth only when compulsory.

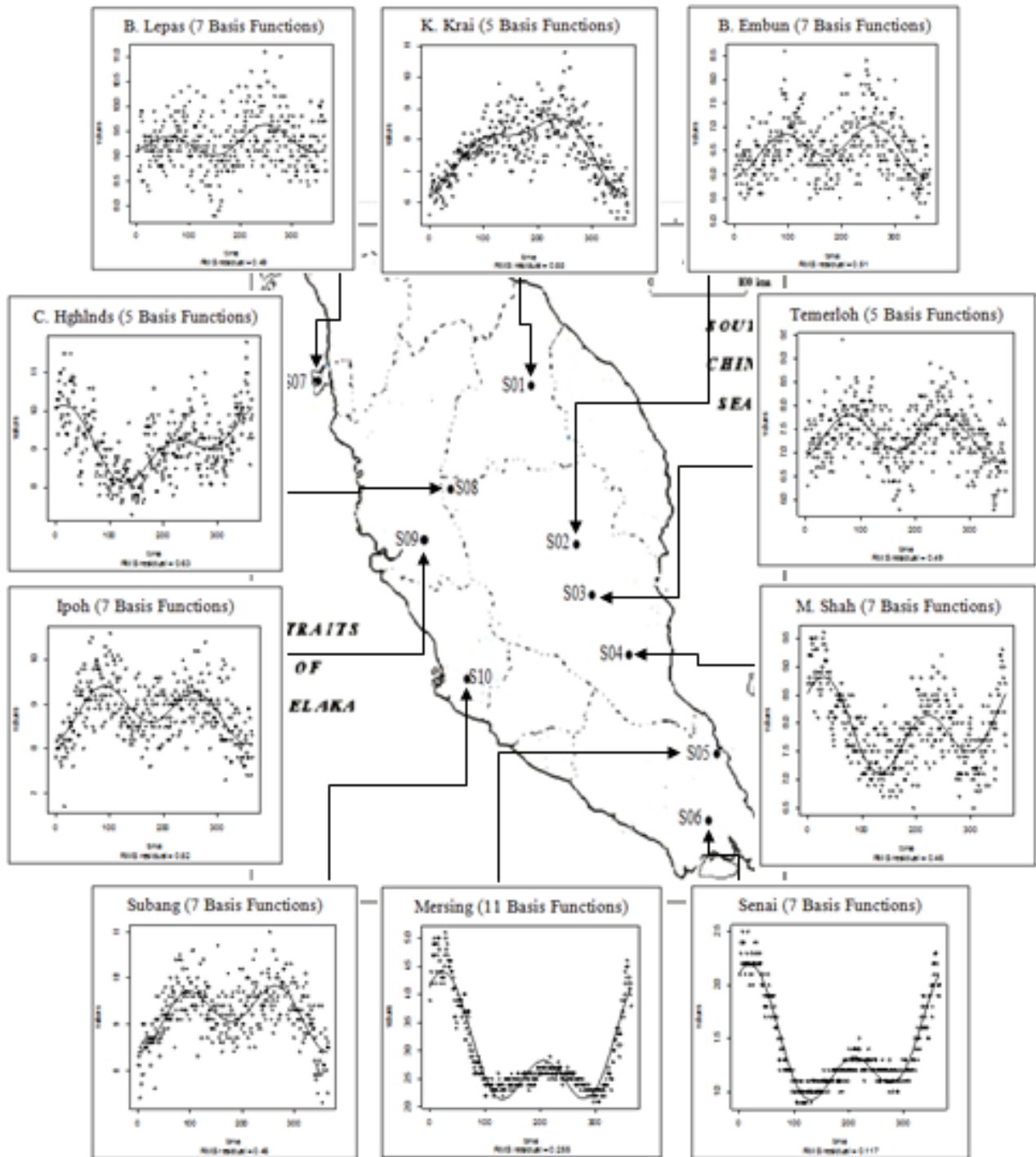


Figure 1 Smoothing wind curves with the required basis functions

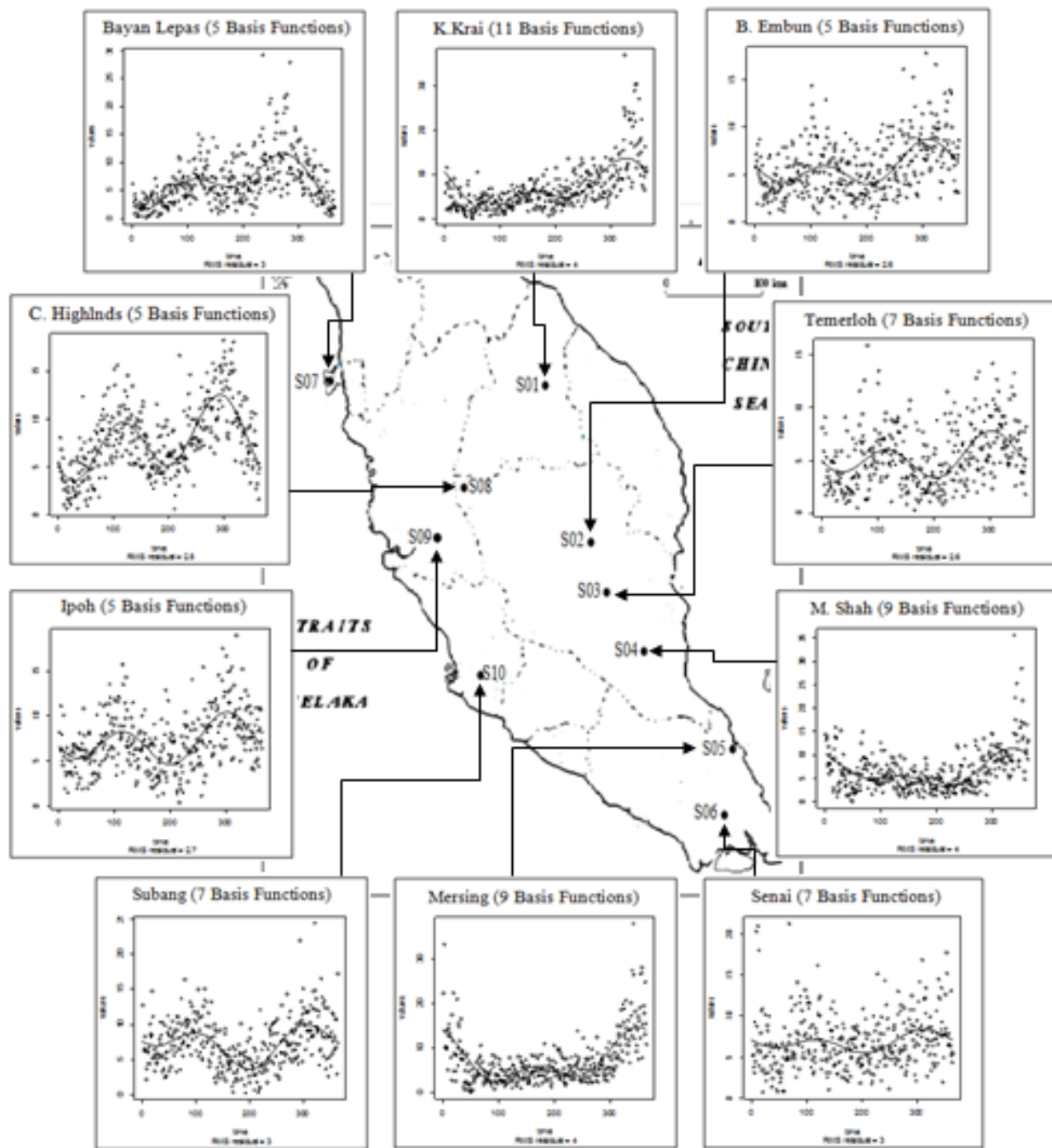


Figure 2 Smoothing rainfall curves with the required basis functions

Table 2 Analysis of deviance of mean daily wind speed and daily rainfall for Batu Embun and Muadzam Shah

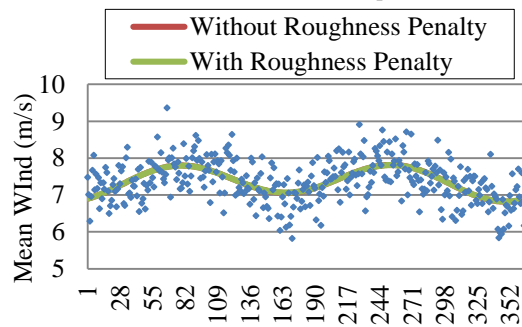
Harmonic	Basis	Degrees of Freedom	Reduction Deviance	Mean Deviance	F	P-Value	Harmonic	Basis	Degrees of Freedom	Reduction Deviance	Mean Deviance	F	P-Value
<i>Batu Embun</i>						<i>Muadzam Shah</i>							
Between Day						Between Day							
1	3	364	134.43	11.53	5.766	23.41	1	3	364	7005.67	908.43	81.30	0.000
2	5	2	27.11	2.128	13.555	55.04	2	5	2	608.61	304.30	27.23	0.000
3	7	2	4.26	1.16	2.128	8.64	3	7	2	348.61	174.31	15.60	0.000
4	9	2	1.16	0.580	2.36	0.31	4	9	2	155.65	77.82	6.96	0.031
5	11	2	1.42	0.710	2.88	0.24	5	11	2	50.81	25.41	2.27	0.321
Residual	13	354	87.19	0.246			Residual	13	354	3955.70	11.17		

Table 3 Values of lambda, degrees of freedom and GCV by using small basis functions

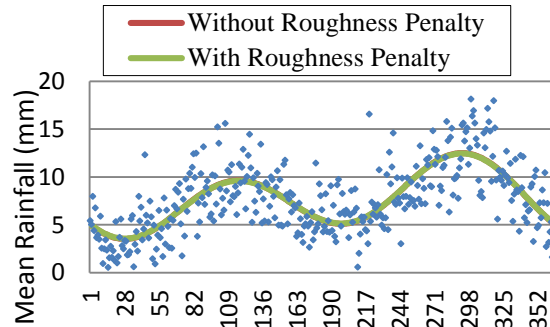
Temerloh (Wind)				C. Highlands (Rainfall)			
Loglam	Lambda, λ	Degrees of Freedom	GCV	Loglam	Lambda, λ	Degrees of Freedom	GCV
5	1.00E+05	4.9998	0.2464	5	1.00E+05	4.9998	6.6052
6	1.00E+06	4.9981	0.2464	6	1.00E+06	4.9981	6.6051
7	1.00E+07	4.9814	0.2463	7	1.00E+07	4.9814	6.6050
8	1.00E+08	4.8287	0.2468	8	1.00E+08	4.8287	6.6405
9	1.00E+09	4.0327	0.2674	9	1.00E+09	4.0327	7.8917

Figure 4 show the smooth functional data for Temerloh and Cameron Highlands with and without roughness penalty by using 365 basis functions. Roughness penalty are added to the least square when constructing the functional data object from observed data in order to reduce the unwanted errors. Table 4 provides the values of sum of squared residuals for both variables. The results indicated that the sum of squared residuals for

smoothing curve with roughness penalty is smaller than the case without roughness penalty. In conclusion, it is better to smooth the data with roughness penalty when using a large number of bases functions.

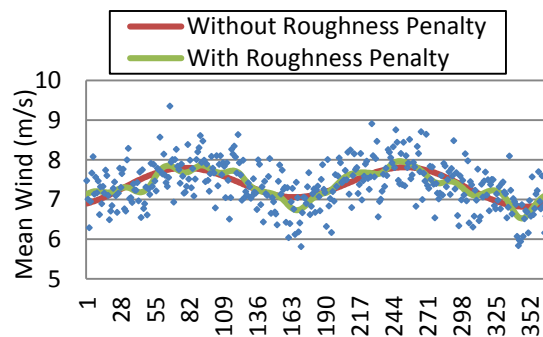


(a)

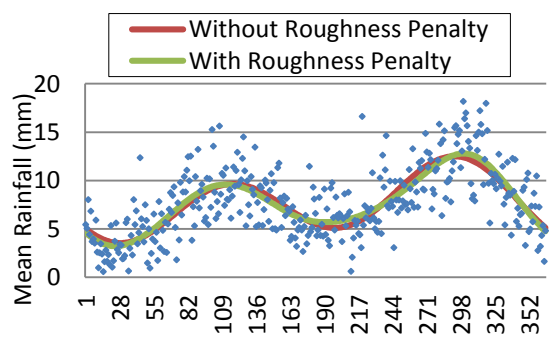


(b)

Figure 3 Smoothing curves with and without roughness penalty for (a) Temerloh (Wind) and (b) Cameron Highlands (Rainfall) using small number of bases functions



(a)



(b)

Figure 4 Smoothing curves with and without roughness penalty for (a) Temerloh (Wind) and (b) Cameron Highlands (Rainfall) using large basis functions

Table 4 Sum of squared residuals for wind and rainfall data for large number of bases functions

Station	Wind		Rainfall	
	Without Roughness	With Roughness	Without Roughness	With Roughness
	Penalty	Penalty	Penalty	Penalty
Kuala Krai	119.34	116.35	4177.18	1781.63
Batu Embun	90.32	76.47	2464.39	1966.58
Temerloh	87.36	75.84	2317.64	2085.38
Muadzam Shah	67.32	47.02	4073.21	2205.96
Mersing	9.93	1.14	4735.20	3600.44
Senai	3.54	1.23	3664.08	2565.81
Bayan Lepas	84.02	82.32	4192.37	4171.56
C.Highlands	143.44	18.00	2346.73	2255.09
Ipoh	92.57	91.94	2641.39	2251.47
Subang	72.21	71.85	3088.01	3065.67

4.3 Summarize the Pattern of Data using The Functional Descriptive Statistics

In this sub-section, we recast the concepts of mean, standard deviation, covariance and correlation into functional terms. Figure 5 shows the average pattern of the mean and standard deviation of wind and rainfall functions for all ten stations in Peninsular Malaysia. High mean speed is observed in August

with 7.3m/s while low mean speed is recorded in June and November. On the other hand, a large variability of wind speed is observed between March and April, and September to October. For rainfall data, high mean rainfall function is observed in November to December with large rainfall variability was also observed during the same period of time. It is expected that high mean rainfall is observed during the northeast monsoon months.

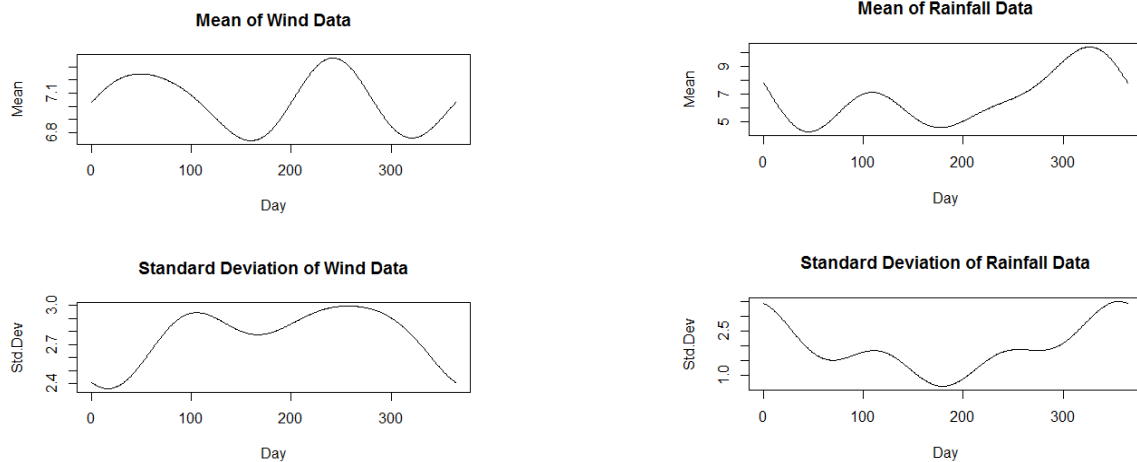


Figure 5 The mean and standard deviations of wind and rainfall functions

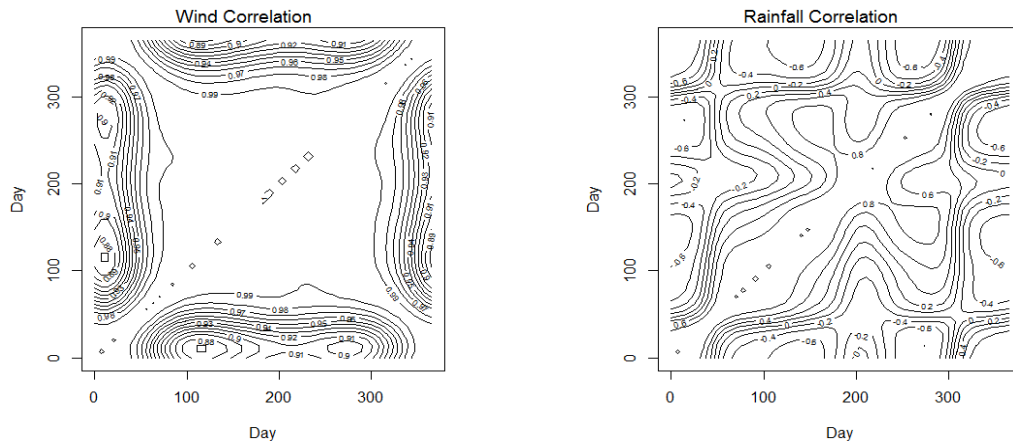


Figure 6 The contour plots of the bivariate correlation functions for wind and rainfall

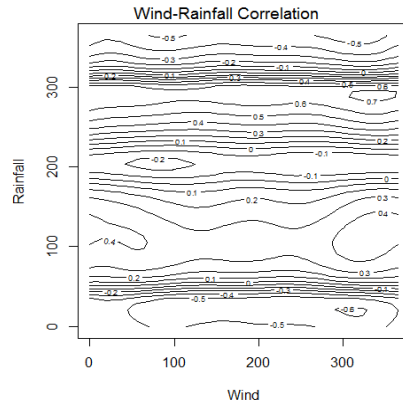


Figure 7 The contour plot of the cross-correlation functions for wind and rainfall

Figure 6 shows a contour plot of the bivariate correlation function $corr_{wind}(t_1, t_2)$ for the wind data and the corresponding plot for rainfall data which is based on 365 days. Generally, the wind data show high positive correlations throughout the year. This concludes that the wind speed is highly related between times at any points of the year. However, different values are recorded for rainfall function over times. It seems that rainfalls are highly correlated during the times in January and February with the correlation value of 0.8. By contrast, moderate negative and positive correlation values are observed from April till October.

The cross-correlation function of two climate variables $corr_{wind,rain}$ is plotted onto Figure 7. The contour plot of cross-correlation functions, t_1 is plotted along the horizontal axis and t_2 along the vertical axis, which represent either wind or rainfall. Generally, it could be said that moderate and low positive and negative correlations are found between two variables throughout the year. Zero values between wind and rainfall functions are observed during the period of July and August within any point throughout the year, which indicate no relationships exist between the variables functions. However, rainfall and wind are highly correlated at the end of the year during the northeast monsoon. It shows the cross-correlation value with 0.70 during November to March gives the strong positive relationship between wind and rainfall, which indicate that the wind speed could influence the rainfall pattern.

5.0 CONCLUSION

This study focused on how to define wind and rainfall data in the form of smoothing curves for ten stations in Peninsular Malaysia. Optimal number of bases functions is determined in describing the characteristics of the wind and rainfall of each station. Based on the smoothed curves obtained for each station, the wind and rainfall patterns are compared. Large variation in rainfall and wind speed of the station required large number of bases functions. Different results are achieved for wind and rainfall curves at the stations may possibly due to several factors such as geographical locations, distance from the sea and monsoons influence.

In establishing the relationship between the two climate variables, it shows a positive correlation during the northeast monsoon season based on the contour plot, while no clear indication of relationships exist for the rest of the months. Therefore, it could be said that the monsoons play a major role

in influencing the relationship between wind and rainfall data.

Several applications of FDA such as functional principal component, canonical analysis, clustering and functional analysis of variance should be employed in future study to examine the variation of climate variables and establish relationship between the climate variables.

Acknowledgement

The author wishes to acknowledge the Malaysian Meteorological Service (MMS) for providing the daily wind and rainfall data for this study. This research was partially supported by MyBrain15 and Research University Grant QJ130000.2526.07H00 from Universiti Teknologi Malaysia.

References

- [1] Cuevas, A., Febrero, M., Fraiman, R. 2003. An Anova Test for Functional Data. *Comput. Stat. Data Anal.* 47: 111–122.
- [2] Ferraty, F., Vieu, P., Viguier-Pla, S. 2006. Factor-based Comparison of Groups of Curve. *Comput. Stat. Data Anal.* 51: 4903–4910.
- [3] Hyndman, R. J., Booth, H. 2008. Stochastic Population Forecasts using Functional Data Models for Mortality, Fertility and Migration. *Int. J. Forecasting.* 24: 323–342.
- [4] Manteiga, W. G., Vieu, P. 2007. Statistical for Functional Data. *Comput. Stat. Data Anal.* 51: 4788–4792.
- [5] Nikitovic, V. 2011. *Functional Data Analysis in Forecasting Serbian Fertility*. Institute of Social Sciences. 2: 73–89.
- [6] Ramsay, J. O., Ramsey, J. B. 2002. Functional Data Analysis of the Dynamics of the Monthly Index of Nondurable Goods Production. *J. Econometrics.* 107: 327–344.
- [7] Ramsay, J. O., Hooker, G., Graves, S. 2005. *Functional Data Analysis with R and Matlab*. Springer. New York.
- [8] Ramsay, J. O., Silverman, B. W. 2005. *Functional Data Analysis*. second ed. Springer, New York.
- [9] Ratcliffe, S. J., Leader, L. R., Heller, G. Z. 2002. *Functional Data Analysis With Application to Periodically Stimulated Foetal Heart Rate Data. I: Functional Regression*. John Wiley & Sons, Ltd. 21. 1103–1114.
- [10] Suhaila, J., Jemain, A. A. 2011. A Comparison of the Rainfall Patterns between Stations on the East and the West coasts of Peninsular Malaysia using the Smoothing Model of Rainfall Amounts. *Meteorol. Appl.* 16(3): 391–401.
- [11] Suhaila, J., Jemain, A. A., Hamdan, M. F., Zin, W. Z. W. 2011. Comparing Rainfall Pattern between Regions in Peninsular Malaysia via a Functional Data Analysis. *J. Hydro.* 411: 197–206.
- [12] Tian, T. S. 2010. Functional Data Analysis in Brain Imaging Studies. *Front Psychol.* 1: 35.