MATEMATIKA, 2018, Special Issue, 167–177

Estimation of Rainfall Curve by using Functional Data Analysis and Ordinary Kriging Approach

[1]Muhammad Fauzee Hamdan*, [2]Abdul Aziz Jemain and [3]Shariffah Suhaila Syed Jamaludin

[1,3]Department of Mathematical Sciences, Faculty of Science
Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia

[2]Faculty of Science and Technology,
Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia

*Corresponding author: mfauzee@utm.my

Article history
Received: 24 September 2018
Received in revised form: 4 December 2018
Accepted: 17 December 2018
Published on line: 31 December 2018

Abstract Rainfall is an interesting phenomenon to investigate since it is directly related to all aspects of life on earth. One of the important studies is to investigate and understand the rainfall patterns that occur throughout the year. To identify the pattern, it requires a rainfall curve to represent daily observation of rainfall received during the year. Functional data analysis methods are capable to convert discrete data intoa function that can represent the rainfall curve and as a result, try to describe the hidden patterns of the rainfall. This study focused on the distribution of daily rainfall amount using functional data analysis. Fourier basis functions are used for periodic rainfall data. Generalized cross-validation showed 123 basis functions were sufficient to describe the pattern of daily rainfall amount. North and west areas of the peninsula show a significant bimodal pattern with the curve decline between two peaks at the mid-year. Meanwhile,the east shows unimodal patterns that reached a peak in the last three months. Southern areas show more uniform trends throughout the year. Finally, the functional spatial method is introduced to overcome the problem of estimating the rainfall curve in the locations with no data recorded. We use a leave one out cross-validation as a verification method to compare between the real curve and the predicted curve. We used coefficient of basis functions to get the predicted curve. It was foundthatthe methods ofspatial prediction can match up with theexistingspatialpredictionmethodsin terms of accuracy,but it isbetterasthe new approach provides a simpler calculation.

Keywords Rainfall amount; functional data analysis; ordinary kriging

Mathematics Subject Classification 62-07, 86A05

1 Introduction

In building a new hydrological structure at a new developmental area, information about nearby rainfall stations are often used to gain an initial overview of the rainfall distribution and pattern. A good town planner, developer or engineer will review and find the potential source of

water supply, estimate the availability on water sources, how many rainfallscan be expected in that area and study the period of drought or wet behaviour. These preliminary steps can more or less, ensure the planned area development projects are built with low risk rate in the hydrology constructions. Therefore, if the pattern of the phenomena throughout the year could be visualized, it can be very useful information in hydrological planning.

One of the methods that can capture the interesting pattern of the data is via functional data analysis. Instead of analysing the discrete data; the functional data analysis method can transform the discrete data into a curve or function. Ramsay and Silverman [1] gave a very good explanation on several functional methods such as principle component analysis, linear model, canonical correlation and discriminant analysis. It has been used in so many applications such as in environmental problem (Gao dan Niemier, [2]) and economy (Laukaitis and Rackauskas, [3]). FDA can also be used to detect the outlier in water quality (Muniz *et al.*, [4]) and Martinez *et al.* [5]on outlier in air quality. Burfield *et al.* [6] used FDA in characterizing the chemical data and conclude that it is a powerful technique to detect the function minima and maxima even though they argued that the computational part was more complex compared to classical multivariate analysis. Sierra et al.[7] and Ruiz-bellido *et al.* [8] shared the same thought that functional data analysis is a promising and valuable tool in their research.Chen *et al.* [9]examined the distribution functions of GDP across different versions of the Penn World Tables and found the need to conduct appropriate analysis to check the robustness of the results. Nowadays, this approach has been widely explored and used in another statistical branch such as in nonparametric statistics (Ferraty and Vieu, [10]), functional analysis of variance (Cuevas et al, [11]) and functional clustering technique (Mizuta, [12]). In this study, we first focus on the exploring of the rainfall pattern using functional data analysis and after that, we try to estimate the rainfall pattern at an unsample location by using ordinary kriging.

2 Rainfall Data

Peninsular Malaysia is situated at the equatorial zone between 1 N and 6 N and longitude from 100 E to 103 E. The seasonal variation of rainfall in Peninsular Malaysia is influenced by two types of monsoon. The first one is the Southwest Monsoon which occurred from May to August. During this period, the whole country experience drier period. The second one is Northeast Monsoon where the east coast and northern area receive heavier rain than other parts of the country. It occurred from November to February. In between these two monsoons are the inter monsoon periods, occurring from March to April and September to October which brings intense convective rain to many areas in Peninsular Malaysia.

Various aspects of rainfall in Peninsular Malaysia have been studied by many researchers. Deni *et al.* [13] studied the occurrence of rainfall events while Suhaila and Jemain [14] studied rainfall intensity. We used the observed daily rainfall data obtained from the Department of Irrigation and Drainage Malaysia. Information on the location of the rainfall data from 75 rain gauge stations located throughout Peninsular Malaysia is available in Table 1.

Table 1: The List of 75 Rainfall Stations with Their Geographical Coordinates

Code	**Station**	**Longitude**	**Latitude**	**Code**	**Station**	**Longitude**	**Latitude**
E1	Aur Gading	101.92	4.35	N1	Kodiang	100.30	6.37
E2	Chebong	102.35	4.12	N2	Kroh	101.00	5.71
E3	Janda Baik	101.86	3.33	N3	Parit Nibong	100.51	5.13
E4	Genting Sempah	101.77	3.37	N4	Pendang	100.48	5.99
E5	Pekan	103.36	3.56	N5	Sik	100.73	5.81
E6	Kuantan	103.22	3.78	N6	Alor Setar	100.40	6.20
E7	Paya Kangsar	102.43	3.90	N7	Ampang Pedu	100.77	6.24
E8	Gua Musang	101.97	4.88	N8	Air Bagan	100.20	5.35
E9	To' Uban	102.14	5.97	N9	Batu Kawan	100.43	5.26
E10	Kota Bharu	102.28	6.17	N10	Bukit Berapit	100.48	5.38
E11	Jambu Bongkok	103.35	4.94	N11	Air Itam	100.27	5.40
E12	Kemasek	103.45	4.43	N12	Bukit Bendera	100.27	5.42
E13	Kampung Jabi	102.56	5.68	N13	Bumbong Lima	100.44	5.56
E14	Setiu	102.95	5.53	N14	Arau	100.27	6.43
E15	Kemaman	103.42	4.23	N15	Guar Nangka	100.28	6.48
E16	Dungun	103.42	4.76	N16	Kaki Bukit	100.21	6.64
E17	Kg. Menerong	103.06	4.94	N17	Kangar	100.19	6.45
E18	Kuala Terengganu	103.13	5.32	N18	Abi Kg. Bahru	100.18	6.51
W1	Lawin	101.06	5.30	S1	Kahang	103.6	2.23
W2	Chui Chak	101.17	4.05	S2	Ladang Kian Hoe	103.27	2.03
W3	Sitiawan	100.70	4.22	S3	Kukup	103.46	1.35
W4	Ladang Boh	101.43	4.45	S4	Rengam	103.35	1.84
W5	Ipoh	101.10	4.57	S5	Johor Bahru	103.75	1.47
W6	Batu Kurau	100.80	4.98	S6	Tampok	103.20	1.63
W7	Alor Pongsu	100.59	5.05	S7	Senai	103.67	1.63
W8	Selama	100.70	5.14	S8	Kota Tinggi	103.72	1.76
W9	Sungai Bernam	101.35	3.70	S9	Batu Pahat	102.98	1.87
W10	Sungai Tua	101.69	3.27	S10	Sembrong	103.05	1.88
W11	Petaling Jaya	101.65	3.1	S11	Separap	102.88	1.92
W12	Subang	101.55	3.12	S12	Kluang	103.32	2.02
W13	Genting Kelang	101.75	3.24	S13	Tangkak	102.57	2.25
W14	Gombak	101.73	3.27	S14	Mersing	103.83	2.45
W15	Sungai Sabaling	102.49	2.85	S15	Endau	103.67	2.59
W16	Sengkang	101.96	2.43	S16	Bukit Asahan	102.55	2.39
W17	Port Dickson	101.80	2.53	S17	Jalan Empat	102.19	2.44
W18	Seremban	101.96	2.74	S18	Merlimau	102.43	2.15
W19	Sg. Lui Halt	102.37	2.08	S19	Melaka	102.25	2.27
W20	Wilayah Persekutuan	101.68	3.16				

3 Functional Data Analysis

Functional data refer to the data in which each observation is curve on some interval where assumed to be smooth. What makes functional data analysis different from other conventional statistical method is the data unit. Many conventional statistical methods treat the numbers or vectors as the units of data. However, functional data analysis uses function or curve as data unit defined on some interval. It also has a wide range of flexibility in the sense that the time observationis not required to be equally spaced for the subject (Song *et al.* [15]).

In FDA, the data under study is the set of random functions, $x_i(t)$. The index $i, i = 1$, ..., n corresponds to the number of stations. Each function x_i is observed only at fixed grid points. Let denote the observation asy_i and we assume that $y_{ij}(t) = x_i(t_j) + e_i$, $j = 1, \ldots, p$where e_i are independent identically distributed random variables. The observation times $t_1, \ldots, t_p$ are equally spaced in time and $p = 365$. To convert the raw observations y_{ij} into functional form, $x_i(t)$ we use Fourier smoothing and this results in the following form

$$x_i(t) = \sum_{k=1}^{K} c_{ik} b_k(t), \tag{1}$$

where $b_k(t)$ denotes the Fourier basis. Although we can use other basis such as B-splines, wavelets or kernel, Fourier basis is found to be more suitable for periodic data. The mean curve can be calculated as follows:

$$\bar{x}(t) = \frac{1}{N} \sum_{i=1}^{N} x_i(t). \tag{2}$$

The mean curve can provide useful information on the pattern of rainfall features over the year.

4 Ordinary Kriging

The first idea of spatial interpolation method was introduced by Krige [16] when dealing with the difficult problem of assessing the ore reserves taken at a few spatial locations. Rapidly the methods spreadinto mainstream statistics at the end of the century (Cressie, [17]). Based on the ordinary kriging, the predicted rainfall amount at an unsampled location $\hat{Z}(\mathbf{s}_0, t)$using measured values $Z(\mathbf{s}_i, t)$is given as:

$$\hat{Z}(\mathbf{s}_0, t) = \sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i, t), \tag{3}$$

where is $\lambda_i$the kriging weight. In kriging, the first step is to examine the data to identify the spatial structure which is represented by empirical semivariogram(Isaaksand Srivastava [18]). A semivariogram is a figure that shows the relationship between semivariance and the distance between all the pairs of available data points as shown in Figure 1

The experimental semivariance is calculated using

$$2\hat{\gamma}(h) = \frac{1}{|N(h)|} \sum_{N(h)} (Z(s_i) - Z(s_j))^2, \tag{4}$$

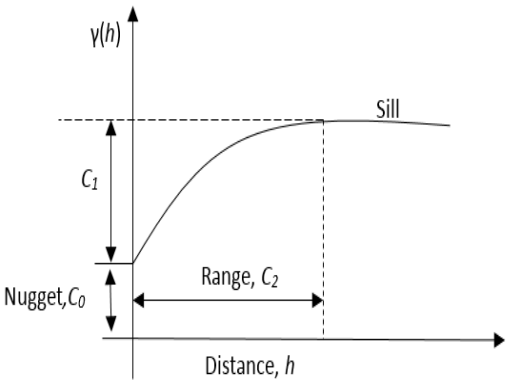

Figure 1: Semivariogram Model

where $\hat{\gamma}(h)$ is the estimated value of the semivariance for a distance of h. $N(h)$ is the number of paired data points at a distance of h. The empirical variogram was fitted to the theoretical variogram function to model the spatial autocorrelation curve. Three types of semivariogram model are spherical model, exponential model and Gaussian model. All three types of semivariogram were tested and the semivariogram that gave the best combination of the parameter C_0, C_1 and C_2 that minimized the error will be selected.

1. Spherical model
$$\gamma(h) = C_0 + C_1 \left[\frac{1.5h}{C_2} - 0.5 \left(\frac{h}{C_2} \right)^3 \right].$$

2. Exponential model
$$\gamma(h) = C_0 + C_1 \left[1 - \exp\left(-\frac{3h}{C_2} \right) \right].$$

3. Gaussian model
$$\gamma(h) = C_0 + C_1 \left[1 - \exp\left(-\frac{h^2}{C_2^2} \right) \right].$$

The "leave out one" cross-validation was used to evaluate how well the model predicted rainfall amount values at the unsampled locations. The indices used during the process was root mean square error (RMSE), where

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{365} \left(x(t) - \hat{x}(t) \right)^2}{365}}. \tag{5}$$

Thus, for each of the sampled locations, there would be one measured value for rainfall amount and the associated predicted value. In this paper, we compared two different approaches to estimate the rainfall amount at an unsampled locations. In the first approach, we implemented

directly the ordinary kriging (DOK) by using the rainfall amount curve obtained from functional data analysis at specific time to estimate the rainfall amount at an unsampled locations. For the second approach, we estimated the coefficient, $\hat{c}\,(\mathrm{s}_0)$at unsampled location by using ordinary kriging (COK) as follow:

$$\hat{c}\,(\mathrm{s}_0) = \sum_{i=1}^{N} \lambda_i \mathbf{c}(\mathbf{s}_i). \tag{6}$$

After that we obtained the rainfall amount curve at unsampled location by using Eq. 1

$$\hat{x}\,(\mathrm{s}_0, t) = \sum_{k=1}^{K} \hat{c}\,(\mathrm{s}_0) b_k(t). \tag{7}$$

5 Results

The spatial variation of rainfall curves across 75 rainfall stations in Peninsular Malaysia has been examined. Figure 2 shows the rainfall curves obtain by using functional data analysis for all the stations. The pattern for rainfall amount curves show the bimodal shaped with many fluctuations. This pattern is believed to be the result of Malaysia's climate which is influenced by the two main monsoon seasons in Malaysia.

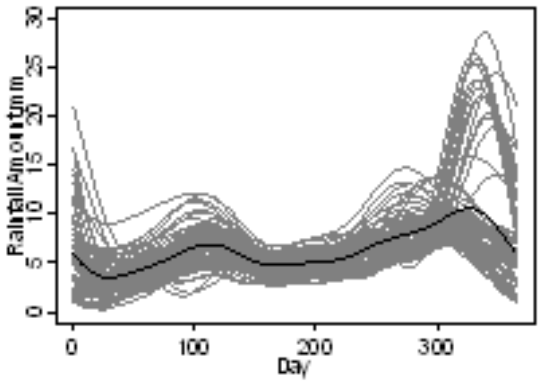

Figure 2: Rainfall Curve for All Stations

The following figures show the rainfall amount curves separated by four regions. This is very useful information to identify the rainfall pattern between regions. North and west areas of the peninsula show a significant bimodal pattern with the curves declining between two peaks at the mid-year. Meanwhile unimodal patterns can be observed in the eastern region which reached a peak in the last three months. Southern areas show more uniform trends throughout the year.

As we have stated earlier, by using ordinary kriging method, we wish to estimate the rainfall amount curve at unsampled location. To compare the result, we appliedthe "leave one out" technique. By using this technique, one rainfall amount curve set will be pulled out (considered as an unsampled location) and the rest of the rainfall amount curves obtained from functional data analysis will be used to predict the rainfall amount curve at the assumed unsampled

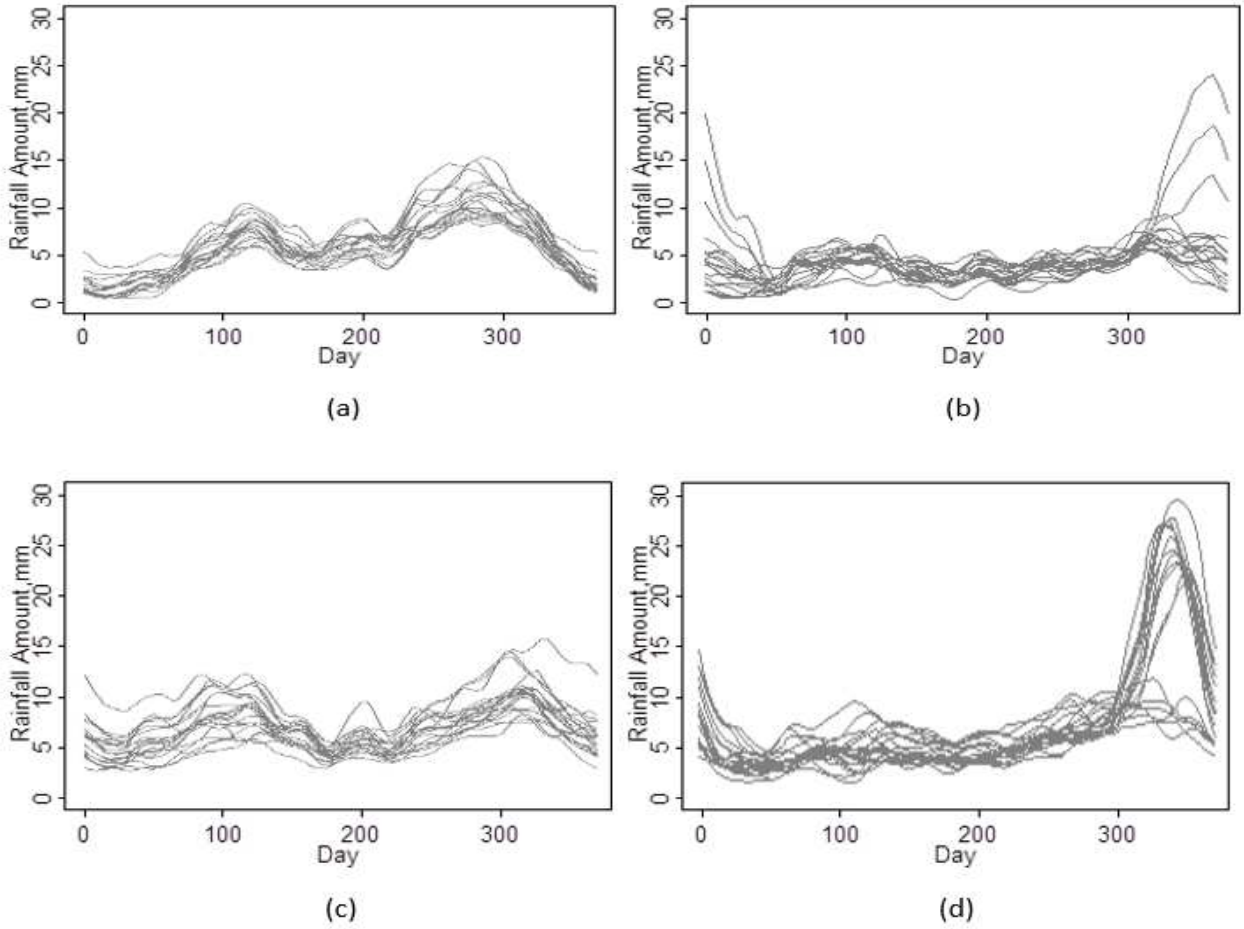

Figure 3: The Rainfall Amount Curve for Each Region (a) Northern, (b) Southern, (c) West and (d) East.

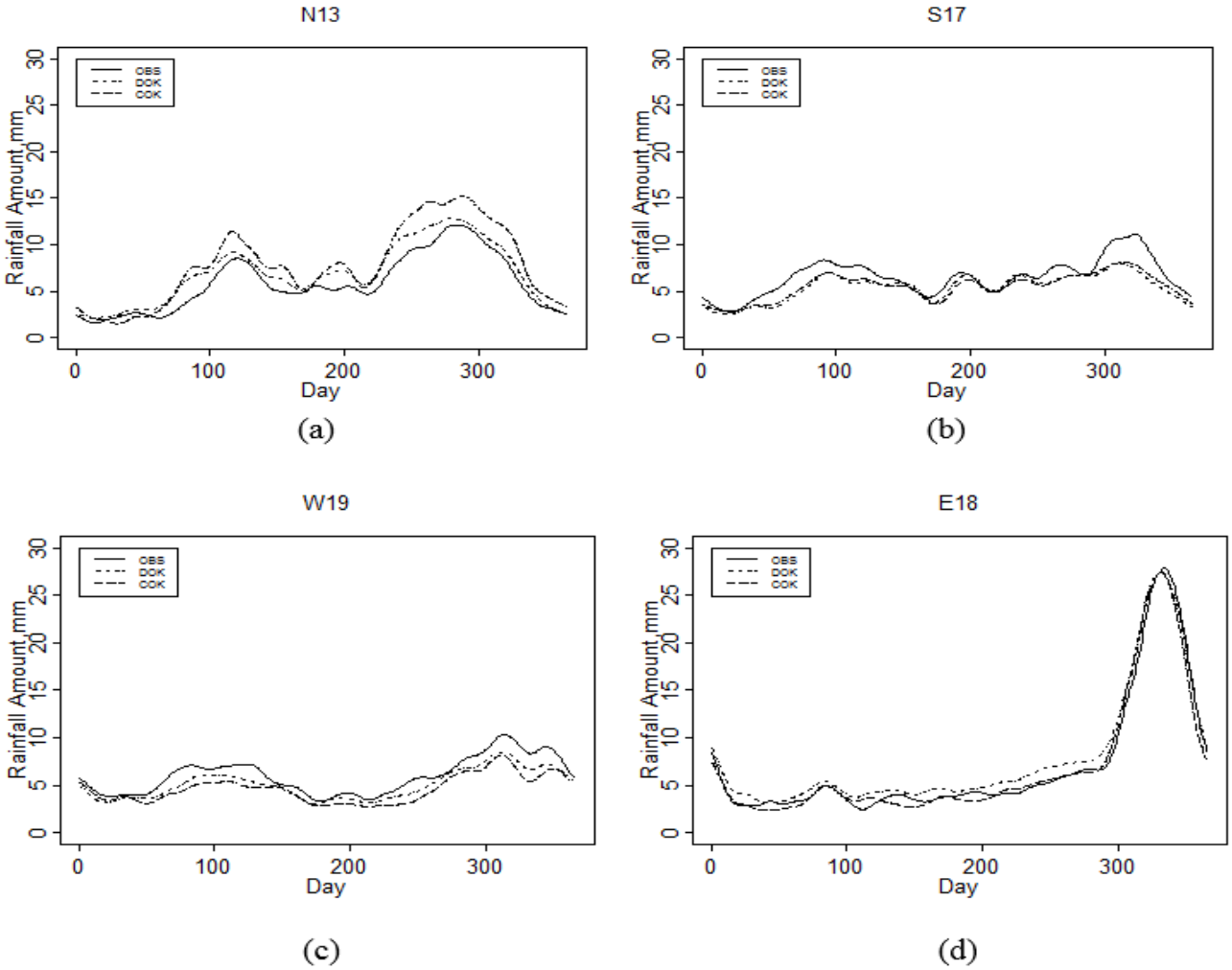

Figure 4: Examples of the Estimated Rainfall Amount Curve Obtained by using DOK and COK Approach for Each Region (a) Northern, (b) Southern, (c) West and (d) East.

location. Figure 4 show the results for both approaches which are DOK and COK. To measure the performance of both approaches, we applied the RMSE and the result is as shown in Table 2.

From the result in Table 2, COK is the best estimator at west and southern regions at 11 and 12 locations respectively. On the other hand, DOK is the best estimation at east region at 10 locations. Both approaches have the same result at nine locations at northern region. In total, COK is the best approach at 40 locations while DOK gives the better estimation of rainfall amount at 35 locations. This result shows that the COK approach can gave better estimation compared to DOK.

6 Conclusion

In this paper we applied the functional data analysis method to obtain rainfall amount curves for 75 rainfall stations in peninsular Malaysia. The coefficients basis obtained from functional

Table 2: Comparison RMSE between DOK and COK for Each Stations

Code	DOK	COK	Code	DOK	COK
E01	**0.6221**	0.6695	N01	0.3701	**0.3644**
E02	1.1139	**0.9707**	N02	0.6432	**0.541**
E03	**2.0666**	2.1117	N03	0.3934	**0.2155**
E04	1.573	**1.5618**	N04	**0.5698**	0.7223
E05	**1.0742**	1.2827	N05	**1.4826**	1.5937
E06	**1.4061**	1.5373	N06	**0.4368**	0.4371
E07	**0.9765**	1.3206	N07	**0.7196**	2.5139
E08	**1.5960**	2.2037	N08	1.0832	**0.9652**
E09	1.4552	**1.3486**	N09	**1.0407**	1.2911
E10	**1.8732**	1.9958	N10	**0.5947**	0.6913
E11	1.2446	**0.9322**	N11	1.3233	**0.9812**
E12	**0.5568**	0.6701	N12	1.572	**1.366**
E13	1.0199	**0.7754**	N13	**1.1523**	2.3198
E14	**1.1155**	1.4796	N14	0.4704	**0.4323**
E15	0.6894	**0.6648**	N15	0.623	**0.5375**
E16	**0.5038**	0.6338	N16	**0.7048**	0.9611
E17	3.3586	**2.6782**	N17	0.4385	**0.3733**
E18	0.9353	**0.7605**	N18	**0.3856**	0.4149
W01	**2.1775**	4.4344	S01	**1.9680**	2.7421
W02	0.3532	**0.1937**	S02	**0.6594**	1.0452
W03	**1.0127**	1.0972	S03	**1.8277**	3.2438
W04	**0.6776**	1.4153	S04	**0.5117**	0.5622
W05	**0.6076**	0.8321	S05	0.9013	**0.8272**
W06	1.7123	**1.3566**	S06	1.7347	**1.7241**
W07	0.8973	**0.5941**	S07	**0.8127**	1.2858
W08	1.6677	**1.1753**	S08	0.8118	**0.7648**
W09	**1.1222**	1.8162	S09	0.8968	**0.788**
W10	**0.4701**	1.2417	S10	**0.6702**	0.93
W11	1.0465	**0.983**	S11	**0.8701**	1.1872
W12	**0.5892**	1.614	S12	1.3131	**1.2048**
W13	0.6922	**0.5247**	S13	1.1405	**0.9241**
W14	0.4462	**0.4204**	S14	4.1574	**3.9187**
W15	1.1320	**1.074**	S15	1.1012	**0.8479**
W16	0.77	**0.5001**	S16	1.1183	**0.866**
W17	1.1521	**0.3561**	S17	1.5205	**1.3653**
W18	**1.0084**	1.5031	S18	1.2509	**1.2063**
W19	**1.1309**	1.6303	S19	0.6878	**0.6176**
W20	0.6593	**0.6494**			

data analysis has been used to predict the rainfall amount at an unsampled location by using ordinary kriging approach. The result shows the estimated curve obtain by using coefficient basis with kriging is better compared the estimated rainfall curve obtain directly by using ordinary kriging.

Acknowledgements

The authors are indebted to the staff of Drainage and Irrigation Department for providing the daily rainfall data for this study. This research was funded by Universiti Teknologi Malaysia with Short Term Research Grant, Vote PY/2017/00322.

References

[1] Ramsay J. O. and Silverman B. W. *Applied Functional Data Analysis: Methods and Case Studies*. Springer. 2007.

[2] Gao H. O., Niemeier D. A. Using functional data analysis of diurnal ozone and NOx cycles to inform transportation emissions control. *Transportation Research Part D: Transport and Environment*. 2008 June 1. 13(4): 221-38.

[3] Laukaitis A, Rakauskas A. Functional data analysis for client's segmentation tasks. *European Journal of Operational Research*. 2005 May 16. 163(1): 210-6.

[4] Muñiz C. D., Nieto P. G., Fernández J. A., Torres J. M., Taboada J. Detection of outliers in water quality monitoring samples using functional data analysis in San Esteban estuary (Northern Spain). *Science of the Total Environment*. 2012 Nov 15. 439: 54-61.

[5] Martínez J., Saavedra Á., García-Nieto P. J., Piñeiro J. I., Iglesias C., Taboada J., Sancho J., Pastor J. Air quality parameters outlier's detection using functional data analysis in the Langreo urban area (Northern Spain). *Applied Mathematics and Computation*. 2014 Aug 15. 241: 1-0.

[6] Burfield R., Neumann C., Saunders C. P. Review and application of functional data analysis to chemical data—The example of the comparison, classification, and database search of forensic ink chromatograms. *Chemometrics and Intelligent Laboratory Systems*. 2015 Dec 15. 149: 97-106.

[7] Sierra C., Flor-Blanco G., Ordoñez C., Flor G., Gallego J. R. Analyzing coastal environments by means of functional data analysis. *Sedimentary Geology*. 2017 July 15. 357: 99-108.

[8] Ruiz-Bellido M. A., Romero-Gil V., García-García P., Rodríguez-Gómez F., Arroyo-López F. N., Garrido-Fernández A. Assessment of table olive fermentation by functional data analysis. *International Journal of Food Microbiology*. 2016 Dec. 5. 238: 1-6.

[9] Chen T., DeJuan J., Tian R. Distributions of GDP across versions of the Penn World Tables: A functional data analysis approach. *Economics Letters*. 2018 Jun 22.

[10] Ferraty F., Vieu P. *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Science and Business Media. 2006.

[11] Cuevas A., Febrero M., Fraiman R. An anova test for functional data. *Computational Statistics and Data Analysis*. 2004 Aug 1. 47(1): 111-22.

[12] Mizuta M., Clustering method for functional data. In *Proceedings in Computational Statistics* Physica-Verlag, A Springer Company. 2004. 1503-1510.

[13] Deni S.M., Jemain A.A., Ibrahim K. Fitting optimum order of Markov chain models for daily rainfall occurrences in Peninsular Malaysia. *Theoretical and Applied Climatology*. 2009 June 1. 97(1-2): 109-21.

[14] Suhaila J, Jemain A. A. 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. *Meteorological Applications*. 2009 Sep. 1. 16(3): 391-401.

[15] Song J. J., Lee H. J., Morris J. S., Kang S. Clustering of time-course gene expression data using functional data analysis. *Computational Biology and Chemistry*. 2007 Aug. 1. 31(4): 265-74.

[16] Krige D. G. A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*. 1951 Dec. 1. 52(5): 119-39.

[17] Cressie N. A. *Statistics for Spatial Data*. John Wiley, 1993

[18] Isaaks E. H., Srivastava, R. M. *An Introduction to Applied Geostatistics*. Oxford University Press. 1989.