Thai Journal of **Math**ematics
Special Issue: The 17th IMT-GT ICMSA 2021
Pages 197–211 (2022)

http://thaijmath.in.cmu.ac.th

ISSN 1686-0209

Clustering Selected Terengganu's Rainfall Stations Based on Persistent Homology

R.U. Gobithaasan[1,*], Zabidi Abu Hasan[1,2], Krithana Devi Selvarajh [1], Khai-Sam Wong[1], Shukri Mamat[3], Mohd Zaharifudin Muhamad Ali[3], Kenjiro T. Miura[4] and Pawe Dotko[5]

[1] *Special Interest Group on Modelling & Data Analytics, Faculty of Ocean Engineering Technology & Informatics, Terengganu, University Malaysia Terengganu. Malaysia.*
e-mail: gr@umt.edu.my (R.U.G)

[2] *Institute of Engineering Mathematics, Faculty of Applied & Human Sciences, University Malaysia Perlis, Perlis, Malaysia.*

[3] *Water Resources Management and Hydrology Division, Malaysia.*

[4] *Graduate School of Engineering, Shizuoka University, Hamamatsu, Japan.*

[5] *Dioscuri Centre in Topological Data Analysis, Warsaw, Poland.*

Abstract Topological Data Analysis (TDA) is an emerging technique rooted from Algebraic Topology that reveals the geometrical structure of high-dimensional data sets. The approach in TDA is twofold; i.e. Persistent homology (PH) which quantifies topological invariants of a given data set, and Mapper which represents the high-dimensional data set into a 1D graph with nodes and edges. In this work, we employ PH as a tool to quantify the first dimensional holes (H_1) in the daily rainfall data set collected between 2012 to 2017 from six rainfall stations located in Terengganu, Malaysia. We divided the rainfall data based on one year (365 days) resulting in each station having five sets of rainfall point clouds. Since a rainfall point cloud consists of 1D data set, direct comparison of rainfalls between stations may not show a clear pattern. Thus, we first embed them into point clouds of 10D with time delay $\tau = 13$, using Takens embedding, preserving its original dynamical state. Next, we employ PH to generate persistence diagram to quantify 1D holes (H_1) in the rainfall point clouds and record its maximum persistence (H_1 lifespan), as its topological feature to characterize the distribution and intensity of rainfall. The first result is; based on past flood events, flood occurred when the year's average persistence score exceeds 13. The second part of this work involves clustering the stations using two approaches; the standard dynamic time warping (DTW) method which matches the rainfall frequency before computing its dissimilarity distance; and the PH approach using five years maximum H_1 lifespan as its distance matrix. The dendrograms produced by both clustering approaches are different, in which DTW has three distinct clusters, but dissimilar to its rainfall distribution. However, PH neatly ranks based on its annual rainfall intensity and recurrence, hence outperforming DTW approach.

MSC: 55-01, 68T10
Keywords: Persistent Homology; Dynamic Time Warping; Clustering,; Time Series; Rainfall

Submission date: 15.03.2022 / Acceptance date: 31.03.2022

*Corresponding author.

Published by The Mathematical Association of Thailand.

1. Introduction

1.1. Background

Malaysia has an equatorial climate with tropical weather of sunshine and rainstorms throughout the year. Malaysian clouds and rainfall prediction is a challenging task due to its tropical weather, receiving sunshine directly and its location being perpendicular to the earths axis of rotation at equator. This situation is worsening due to the impact of climate change.

Some parts of Malaysia are affected by monsoon winds at different intervals of the year and these influence the rainfall at those areas. There are two monsoon winds; the Southwest and Northeast monsoon. The Southwest(SW) Monsoon brings rainfall to the western side of Peninsular Malaysia which occurs between May till September. During these period, the affected areas such as Kuala Lumpur, Penang and Langkawi are prone to floods. Conversely, the Northeast (NE) Monsoon starts from November and lasts till March, bringing heavy rainfall to areas on the east side of Peninsular Malaysia, such as Kelantan, Terengganu, Pahang and Borneo (Sabah and Sarawak). As this monsoon wind is particularly strong, it often brings heavy rain causing devastating floods events [1, 2].

In order to understand the rainfall pattern, researchers investigate the rainfall distributions; comparing between rainfall stations throughout an investigated region. Example of work includes Walsh [3] who clustered rainfall stations in Sabah into four groups (Sandakan, Kota Kinabalu (KK), Tawau and Danum Valley). Even though KK is located near to Sandakan, its rainfall distribution is totally different. Walsh reported that this is due to the nature of KK's shoreline which is parallel to the NE monsoon. Hence, whichever station located at the shoreline perpendicular to the NE monsoon accumulates heavy downpour, whereas those parallel escape the spell. The rainfall distribution in Sandakan is similar to Terengganu due to its exposure to NE monsoon. In fact, the shoreline of Terengganu, Kelantan and Pahang are perpendicular to NE monsoon thus exposed to intense rainstorm during this period.

Another reason of rainfall variation could be due to its topography; whether it is located in a valley or blocked by hills. It is obvious that KK is protected from NE monsoon, due to the location of KK mountains which are parallel to NE monsoon, hence, creating a pathway to flow through. Sandakan however, has an inverse effect in which the mountains are blocking the NE monsoon thus accumulating enough precipitation for a heavy downpour. It is worth noting that Terengganu has the same contour to Sandakan, where the inner parts of Terengganu are hilly. Table (1) shows major floods that occurred between 2012 –2017 in Terengganu.

Clustering rainfall time series may aid in understanding the characteristics of the station and its surrounding area, giving us the idea on flood formation dynamics. Visual graphics can be difficult to differentiate rainfall data between stations. We may carry out statistical tests, for example, homogeneity tests, to check if the variance between stations are similar[23]. A better way to differentiate the characteristics between stations is hierarchical agglomerative clustering (HAC) method which reveals the linkages between stations. In this paper we employed two types of HAC; Complete-linkage (farthest neighbour) and Single-linkage (nearest neighbour) [18, 19]. Computing HAC requires the choice of distance measure which represents similarity/dissimilarity between stations. Dynamic Time Warping (DTW) is the gold standard to compare time series data [20–22]. It maps

TABLE 1. Terengganu flood details extracted from [4]

Year	District	Affected Victims	Victims Killed	Financial Loss
2013	Kemaman	7780	4	US$ 116 million
2014	Kemaman, Dungun, Hulu T'ganu, Kuala Terangganu, Besut	83736	8	US$ 231 million
2017	Kemaman, Dungun, Hulu T'ganu, Kuala Terangganu Besut	12910	0	US$ 1.25 billion

data points between a pair of time series by minimizing the pairwise Euclidean distance between them.

Yet another way to investigate rainfall time series is by investigating the topological information of time series. Prior attempt to investigate topological information of time series with embedding dates back to 1993 where Muldoon et al. [13] classified periodicity based on Betti numbers. However, it did not attract much attention. In 2012, Silva et al.[9] first proved that if the time series is periodic, then it corresponds to a circle in phase space. Similarly, a quasi-periodic dataset which consists of the composition of periodic sets corresponds to d-dimensional torus embedded in d-dimensional phase space. These topological information can be quantified by Persistence Diagram (Dgm) obtained from Persistent Homology's (PH) algorithm.

In 2015, Perea & Harer [14] investigated the geometric structure of time series using truncated Fourier series and stated that a periodic time series has the form of an elliptic shape thus worked on relating its semi-major and semi-minor axes to periodicity and the window size of the embedding. They reported that maximum persistence (the prominent topological feature in Dgm) as an effective measure of periodicity after centering and normalizing the dataset. In the same year, Pereira & Mello [15], proposed the application of PH and used Dgm's summaries to cluster time series and spatial data. In this work, we used dissimilarity distance extracted from Dgm to generate HAC to cluster six rainfall stations located in Terengganu, and compared the results with the classical DTW approach.

The objective of this paper is twofold; (1) to identify the threshold for flood events due to the Northeast monsoon using Persistent Homology (PH), and (2) to cluster six rainfall stations. In order to achieve the second objective, we compared the outcome using Dynamic Time Warping (DTW) and PH. DTW is the gold standard to compare time series, whereas PH is an emerging technique on quantifying the shape of data by means of topology. Hence, the comparison of these two methods would be useful in validating the overall rainfall pattern in Terengganu.

The rest of the sections are arranged as follows: the next section introduces the rainfall dataset and its details. It is followed by the methodologies. Section 3 describes the application of DTW and PH for clustering rainfall stations and its detailed discussion, and finally Section 4 concludes this work with future research directions.

1.2. Rainfall Data

There are over 90 rainfall stations located in Terengganu, where they are numbered based on their location and maintained by the Department of Irrigation and Drainage,

Malaysia. Some of them have been functioning since the 1940s, whereas a number of them have been closed/relocated. Due to these circumstances, there exist missing data for most of the stations. Based on the years of flood/droughts event and data availability, we have chosen six rainfall stations located in various district in Terengganu as depicted in Table (2). The rainfall data period is selected from June 2012 to May 2017 where there is no missing value.

TABLE 2. Details of selected stations for the study.

District	**Station ID**	**Station Name**	**Latitude**	**Longitude**
Kemaman (K)	4234109	JPS. Kemaman	$04°13'55''$	$103°25'20''$
Dungun(D)	4734079	Sek. Men. Sultan Omar Dungun	$04°45'45''$	$103°25'10''$
Hulu Terengganu (H.T)	5029036	Rumah Pam Paya Kemat	$05°00'30''$	$102°58'10''$
Marang (M)	5131064	Sek. Men. Bkt. Sawa	$05°11'30''$	$103°06'00''$
Kuala Terengganu (K.T)	5331048	JPS. Kuala Terengganu	$05°19'05''$	$103°08'00''$
Besut (B)	5625003	Paya Peda Besut	$05°36'00''$	$102°30'55''$

In this paper, we defined a year period as; starting from 1^{st} June to 31^{st} May of the following year. By doing so, we are able to capture each cycle of the NE monsoon season that occurs between November to March the following year. Table (3) depicts annual rainfall ordered from the highest (top) to lowest (below). The station in bold represents Paya Peda Besut situated in Besut (station ID 5625003) having the highest annual rainfall and JPS. Kemaman (station ID 4234109) having the lowest rainfall among these six stations. Readers may obtain the data by contacting the Department of Irrigation and Drainage, Malaysia.

TABLE 3. Annual rainfall from highest(top) to lowest(bottom) [5]

Station ID	**2012-2013**	**2013-2014**	**2014-2015**	**2015-2016**	**2016-2017**
5625003	5577.9	3681.6	4921.0	3050.0	5330.1
5029036	4791.9	3798.1	4391.4	2896.5	4323.2
5131064	3571.6	2702.7	3627.0	2086.3	4001.0
4734079	2787.0	2792.8	3204.2	1853.3	3424.5
5331048	2881.2	2028.9	2470.5	1803.8	3411.4
4234109	2675.3	2899.8	2713.8	1661.5	3236.3

Line graphs in Figure (1) illustrate daily rainfall distribution of six stations . They are colour coded based on the definition of annual rainfall as stated above. Obviously one would not be able to differentiate rainfall from this plot. Since the rainfall data possess non-normal distribution, we employed Flinger-Killeen and Bartlett test. Flinger-Killen test indicated that p-values are 0.000 for all the stations, hence indicating the variances are heterogeneous. The results of Bartlett test as shown in Table (4) are similar except for station JPS Kemaman and Sek. Men. Sultan Omar Dungun, where the p-value more

FIGURE 1. Daily rainfall distribution for six stations between June 2012 May 2017.

than 0.05, suggests that they have equal variances. In order to verify these results, we need to carry of hierarchical agglomerative clustering (HAC) method to find linkages between stations.

TABLE 4. Bartlet Homogeneity Tests

	4234109	**4734079**	**5029036**	**5131064**	**5331048**	**5625003**
4234109						
4734079	0.0764*					
5029036	0.0000	0.0000				
5131064	0.0018	0.0000	0.0000			
5331048	0.0000	0.0000	0.0000	0.0000		
5625003	0.0000	0.0000	0.0000	0.0000	0.0000	

2. Methodology: Persistent Homology

Topological Data Analysis (TDA) is a new method developed with the basis of Algebraic Topology to quantify the structure of high-dimensional datasets [7]. Recently, it has been used as an effective data mining and exploratory data analysis in various fields. Understanding the global structure and overall connectivity of dataset which is in the form of n-dimension point cloud is an essential initial task to reveal the linearity, distribution, clusters and groups abnormality. TDA fits this job well as it has the capability by formalizing the datasets connectivity using simplicial complexes with the aid of metric spaces. Persistent homology [6, 7] is one of the most used methodologies in TDA. By increasing the radius (ϵ) of a cover (usually a ball) between data points in the point cloud and continuously building simplicial complex K, the persistence of the n-dimension holes (H_n) is traced by pairs of births and deaths (b_i, d_i). We call this process as the filtration of the point cloud which encodes the birth and death of H_n homological features. The output is a multi-set consists of pairs of (b_i, d_i) plotted either in the form of barcodes or Persistence Diagrams (Dgm). The invariant that persists over ϵ increment is denoted as Betti number β_n is the indeed the cardinality of H_n quantifies n-dimension holes, starting from β_0 as path-connected components, β_1 as loops, β_2 as voids etc. Lifespan or the persistence of a point x_i in Dgm is denoted as $L(x_i) = d_i - b_i$, thus the pair which has the maximum lifespan is the most prominent feature and we denote it $L^m(Dgm) = \underset{x_i \in Dgm}{max}\, L(x_i)$. Readers are referred to [6, 24] for a detailed explanation on Persistent Homology.

In this work, we employed TDA tool available in Python called Giotto-TDA [8] to generate Dgm. It is compatible with scikit-learn allowing TDA as part of large-scale Machine Learning tasks.Using Giotta-TDA, we constructed Vietoris-Rips complex of the rainfall point cloud with coefficient field, $\mathbb{F} = 11$ to obtain its topological summaries in the form of Dgm.

2.1. Takens' Embedding

Rainfall time series are indexed in time order or can be defined as a sequence taken at successive equally or uneven spaced points in time. Since time series are usually 1D signals, topologically it can be trivial. One way to realize the topology of time series is by embedding them in a higher dimension using Taken's embedding [10] as follows:

Definition 2.1. Given a time-series $f : t \to R$, and a parameter τ, a time-delay embedding is a lift to a time-series $\phi : t \to R^d$ defined by

$$\phi(t) = (f(t), f(t+\tau), \ldots, f(t+(d-1)\tau)) \tag{2.1}$$

where $\tau \times d$ is the window size. The result of 1D time series are now in the form of R^d point cloud, which can be considered as a manifold.

Theorem 2.2. *Let M be compact manifold of dimension m. For pairs (ϕ, y) with $\phi \in Diff^2(M)$, $y \in C^2(M, R)$, it is a generic property that the map $\Phi_{(\phi,y)} : M \to R^{2m+1}$ defined by*

$$\Phi(x)(\phi, y) = (y(x), y(\phi(x)), ..., y(\phi^{2m}(s)) \tag{2.2}$$

is an embedding.

Taken's theorem states that almost every 1D time-delay embedding can recover the underlying manifold and the dynamics of the system. Hence, the choice of time-delay parameter (τ) and dimension (d), are important in point cloud reconstruction to unfold the attractor. An example of failed reconstruction includes when some points falsely appear to be its neighbour in the embedding. There are two ways to identify optimal τ and d in order to successfully preserve the dynamics of time series; mutual information [11] and false nearest neighbours [12]. Both of these methods are readily available in Giotto-TDA, hence making our tasks to search for optimal parameters easier.

3. Application

3.1. Rainfall point cloud and its Persistence Diagram

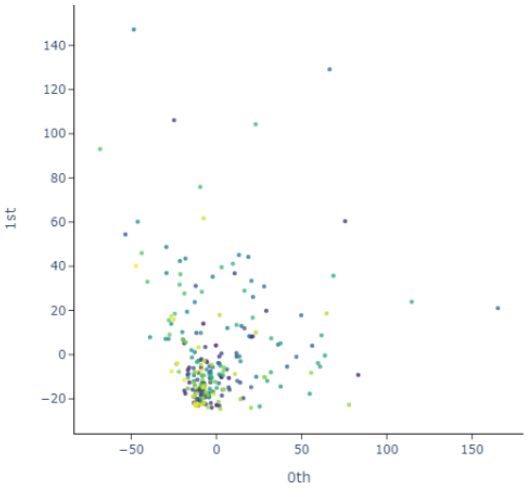
The first step is to identify the optimum embedding parameters for the rainfall stations using mutual information and nearest neighbour. Table (5) shows the corresponding d and τ for each rainfall station. On average, we obtained $d = 10$ and $\tau = 13$, where the window size is 130 days. The smallest annual window embedding is $8 \times 13 = 104$ days which was computed for 2013 - 2014 during one of worst flood event in Malaysian history.

Table 5. Optimum parameters for Taken's embedding using mutual information and nearest neighbour.

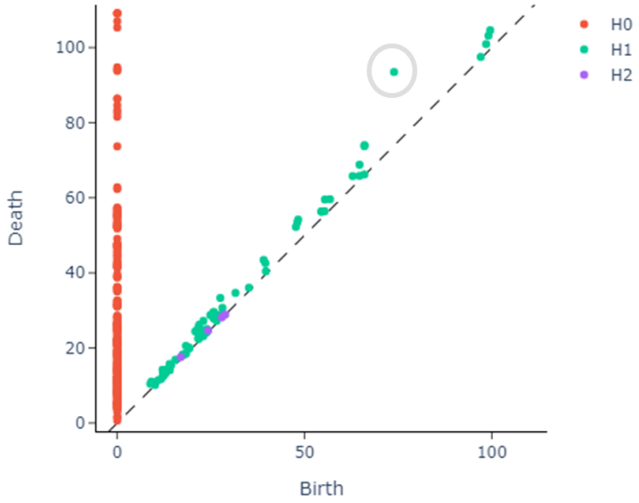
	2012 - 2013		2013 - 2014		2014 - 2015		2015 - 2016		2016 - 2017	
Station ID	d	τ	d	τ	d	τ	d	τ	d	τ
4234109(K)	9	13	10	9	11	18	14	12	12	10
4734079(D)	9	17	9	15	8	16	8	15	13	9
5029036(H.T)	10	16	2	19	8	17	12	5	16	8
5131064(M)	11	17	4	7	11	17	9	11	6	11
5331048(K.T)	9	13	10	19	11	18	14	12	12	10
5625003(B)	9	16	10	11	11	18	14	7	15	11
Average:	10	15	8	13	10	17	12	10	12	10

Using the computed average parameter embedding, we generated persistence diagrams for each rainfall station. Figure (2a) is an example of a point cloud for rainfall station Sek. Men. Sultan Omar (4734079) situated in Dungun Terengganu reduced to 2D using Principal Component Analysis (PCA) where the x-axis and y-axis representing its first and

second principal components. Its corresponding persistence diagram is depicted in Figure (2b). The circled point in the persistence diagram is furthest away from the diagonal line indicating the most persistent loop H_1, thus producing the maximum lifespan: $L^m = 14.6$. Figure (3) shows the lifespan pairs (b_i, d_i) of six rainfall stations during the devastating flood that occurred between June 2012 May 2013. Take note that the station in Besut called Paya Peda Besut has the highest L^m among the rest of the stations.



(A) Point cloud with the dimension of 248×10 reduced to 2D using PCA for visualization



(B) Persistence diagram

FIGURE 2. Embedded point cloud and its corresponding persistence diagram for Sek. Men. Sultan Omar Dungun between June 2012 May 2013.

Table (6) shows the 1-dimensional topological information consists of L_{max} and the number of loops ($\#H_1$) for six stations investigated in the paper based on the location of their district.

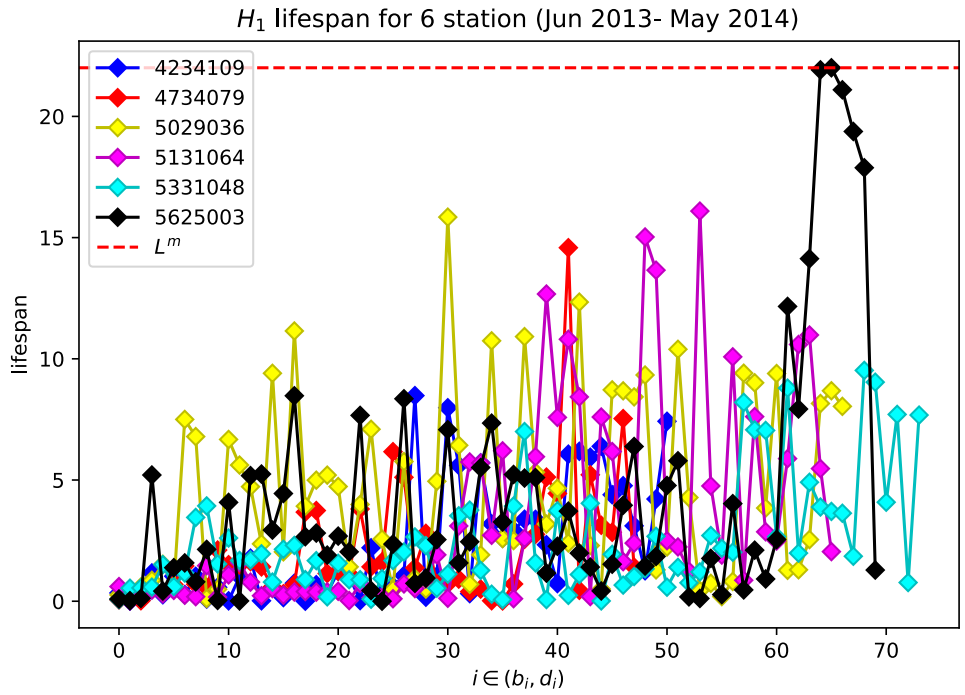


FIGURE 3. Comparison of H_1 Lifespan for six stations.

TABLE 6. L^m and number of H_1 pairs ($\#H_1$) for six station between June 2012 to May 2017.

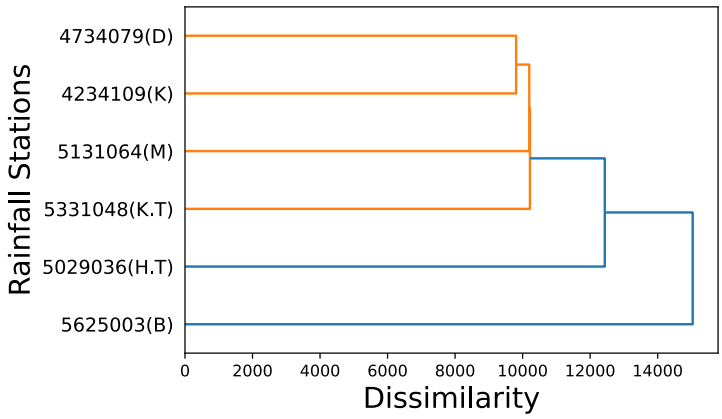
Year		**K**	**D**	**H.T**	**M**	**K.T**	**B**	**Average**
2012-2013	L_{max}	6.0	19.6	16.1	13.0	7.3	17.3	**13.2**
	$\#H_1$	103	76	114	90	46	42	79
2013-2014	L_{max}	8.5	14.6	15.8	16.1	9.5	**22.0**	**14.4**
	$\#H_1$	51	48	67	66	74	70	63
2014-2015	L_{max}	8.5	18.0	10.0	15.2	11.3	5.9	11.5
	$\#H_1$	51	55	52	51	42	42	49
2015-2016	L_{max}	9.3	10.0	5.5	6.8	8.2	6.8	7.8
	$\#H_1$	69	45	90	83	60	82	72
2016-2017	L_{max}	14.7	11.7	12.3	9.5	12.4	18.1	**13.1**
	$\#H_1$	75	88	99	92	94	69	86

Based on this table, it is apparent that we would be able to relate L^m to flood events (Table 1), where the average of L^m is $\overline{L^m} > 13$ for 2012-2013, 2013-2014 and 2016-2017. For year 2015-2016, Malaysia experienced 2015/2016 El-Nino drought event [16], thus, $\overline{L^m}$ was the lowest, amounting to $\overline{L^m} = 7.8$. However, the number of pairs of H_1 does not reflect the intensity of the rainfall, thus fails to reflect the flood events. Hence, we can set $\overline{L^m} > 13$ as a good threshold value for the probability of flood occurrence. In the context of topological indicator, the higher L^m, the bigger H_1 loop exist in the rainfall dataset, hence indicating periodicity. From Table (6), it is apparent that station Paya Peda Besut (5625003) has the highest score of $L^m = 22$ during 2013-2014, indicating higher amplitude of periodic rainfall in this year.

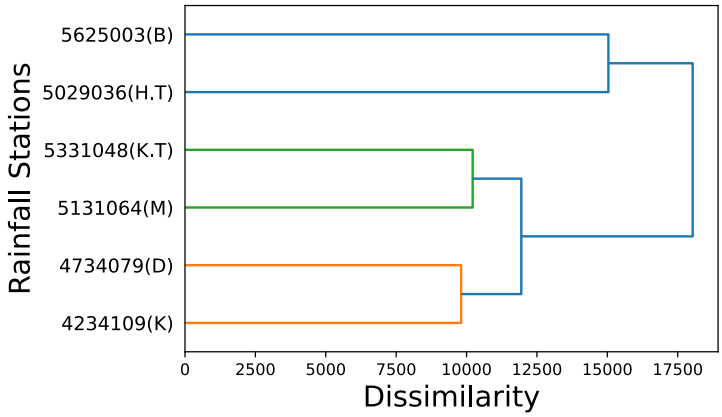
3.2. Discussion: Clustering Rainfall Stations

In this paper, we generate two types of clusters based on two dissimilarity matrix obtained from DTW [17] and L^m. We employed Cophenetic Correlation Coefficient (CCC) to measure the accuracy of clusters based on the obtained dissimilarity matrix; $CCC \approx 1$ which indicates that the resulted dendrogram with its dissimilarity matrix approximates well in the cluster space. A standard acceptable dendrogram has $CCC > 0.7$.

DTW calculates the optimal matches between each of the six rainfall frequencies, hence producing DTW dissimilarity matrix, which is used to generate a dendrogram. For the five years rainfall dataset, Single-linkage produced $CCC = 0.89$ and Complete-linkage produced $CCC = 0.85$. Both has almost similar clusters as shown in Figure (4), where Complete-linkage cluster clearly highlights the sub-cluster better than Single-linkage cluster which groups based on either coastal or inland topography. Its corresponding map is shown in Figure (5).



(A) Sinlge-Linkage DTW clustering Approach.



(B) Complete-Linkage DTW clustering Approach.

Figure 4. Clustering rainfall stations based on Dynamic Time Warping.

The dendrograms generated using L^m as dissimilarity matrix, produced different results where Single-linkage produced $CCC = 0.81$ and Complete-linkage produced $CCC = 0.84$ (Figure 6). For Complete-linkage, with the phenon line at 17, the first cluster in orange consists of two stations near shoreline, followed by next two clusters in between shoreline

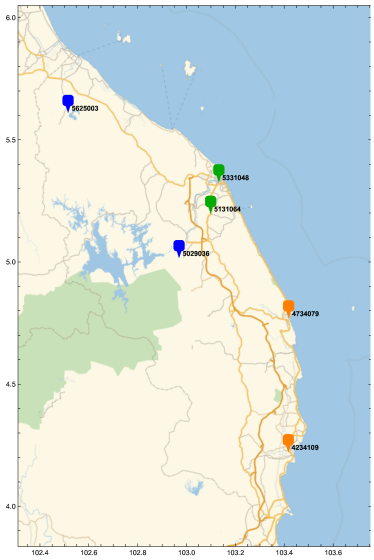


FIGURE 5. Stations coloured based clusters produced by DTW's Complete-linkage approach.

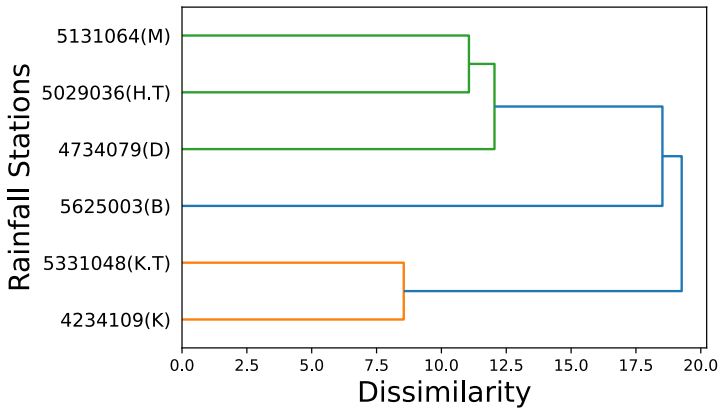
and interior, and finally last stations in the interior part of Besut, Terengganu. Based on corresponding map is shown in Figure (5), it is not clear wheather DTW clusters based on the location in the map.

In order to understand the difference of clusters obtained from DTW and PH, we first compare the outcome of PH with total annual rainfall. The way PH clusters tend to be similar to the orders of total annual rainfall depicted in Table (3). This indicates L^m measures the intensity of rainfall in which DTW fails to detect. Hence, it is now apparent that the reason behind station JPS Kuala Terengganu (5331048) and station JPS Kemaman (4234109); both has smallest annual rainfall as compared to the rest of the stations.

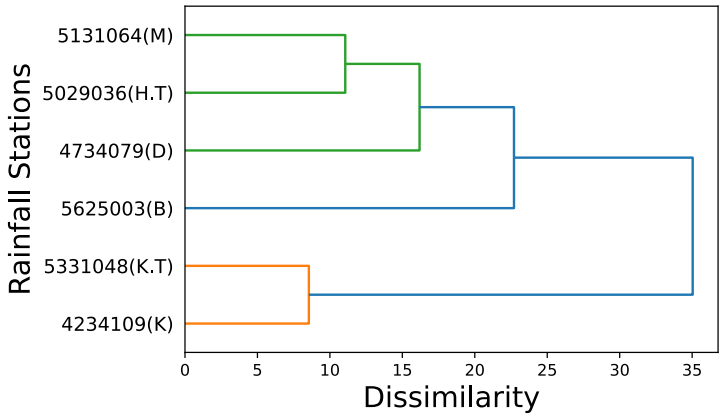
In the section earlier Bartlett test showed that station JPS Kemaman (4234109) is homogeneous with station Sek. Men Sultan Omar Dungun between, is reflected by DTW clusters, not by PH or the boxplots of the rainfall.

4. Conclusion AND Future Work

This work addresses the investigation of rainfall distribution of six stations located at different districts in Terengganu for the period of June 2012 - May 2017. Two objectives of this work have been answered in detail where the first part of this work involves quantifying the maximum score of H_1 loops, denoted as L^m after embedding the rainfall dataset in 10D. The threshold of $\overline{L^m} > 13$ clearly shows the occurrence of flood events in this period. The second part of this work involves clustering these stations based on DTW and L^m using HAC method. The dendrogram produced using L^m clearly outperforms the dendrogram produced by DTW approach, hence, elucidates the relationship of the stations based on rainfall's distribution and its intensity. It is anticipated that L^m can be used as a prominent topological feature for clustering rainfall stations. Additionally, it can also be used for stations with missing values and, can be implemented as labels for machine learning techniques.



(A) Single-Linkage



(B) Complete-linkage approach

FIGURE 6. Clusters based on L^m using Single and Complete linkage approach.

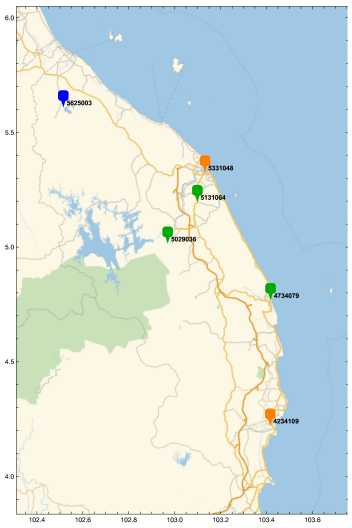
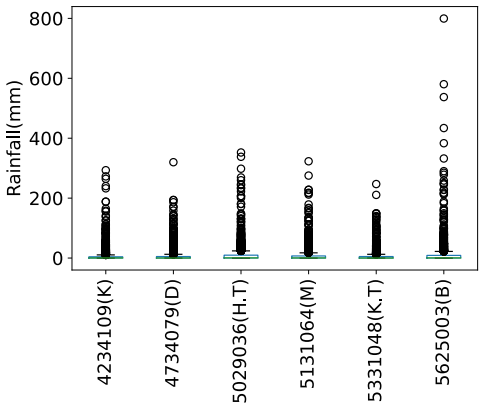
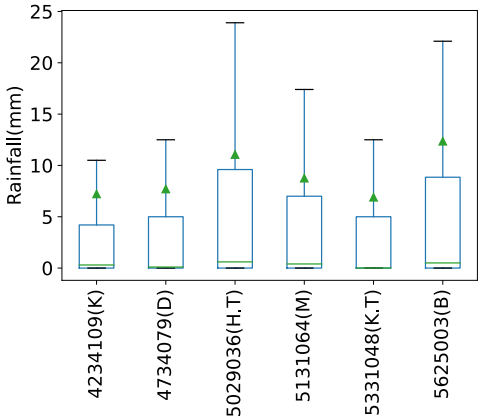


FIGURE 7. Stations coloured based clusters from L^m using Complete-linkage approach.



(A) With outliers.



(B) Without outliers.

FIGURE 8. Boxplot of rainfall with and without outliers where the green markers indicate mean values.

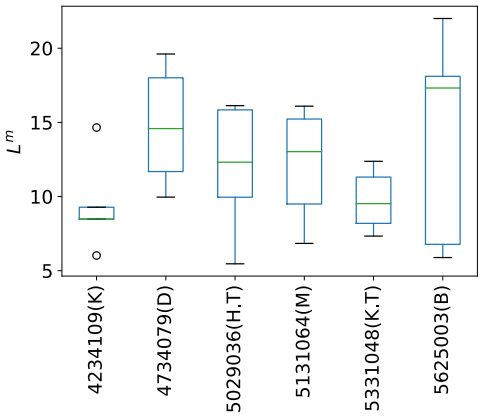


FIGURE 9. Boxplot of L^m distribution based on Table (6).

Future work includes rainfall clustering based on daily rate of change by investigating the sliding-window of rainfall, rather than using dataset segregated annually. We are also

planning to use persistence landscape and distance-to-measure filtration for clustering rainfall dataset with high missing values.

Acknowledgements

This research was supported by the Ministry of Education (MOE) Malaysia through Fundamental Research Grant Scheme (FRGS/1/2019/STG06/UMT/02/2). We would like to thank the Dept. of Irrigation & Drainage Malaysia for providing rainfall dataset which was utilised in this work.

References

[1] N.W. Chan, Impacts of Disasters and Disaster Risk Management in Malaysia: The Case of Floods. Resilience and Recovery in Asian Disasters. (2014) 239265.

[2] F. S. Buslima, R.C. Omar, T. A. Jamaluddin, H. Taha, Flood and Flash Flood Geo-Hazards in Malaysia. International Journal of Engineering & Technology. 7(4.35) (2018) 760–764.

[3] R. P. D. Walsh, Drought frequency changes in Sabah and adjacent parts of northern Borneo since the late nineteenth century and possible implications for tropical rain forest dynamics. Journal of Tropical Ecology. 12 (1996) 385–407.

[4] K.Y. Lim, N.A. Zakaria, K.Y. Foo, A shared vision on the historical flood events in Malaysia: Integrated assessment of water quality and microbial variability, Disaster Adv. 12 (2019) 11-20.

[5] Department of Irrigation and Drainage Malaysia (DID), Yearly Rainfall Data, Yearly Annual Reports.

[6] A. Zomorodian, G. Carlsson, Computing persistent homology. Discrete Comput. Geom. 33(2) (2005) 249–274.

[7] G. Carlsson, Topology and data, Bull. Amer. Math. Soc. 46 (2009) 255–308.

[8] G. Tauzin, U. Lupo, L. Tunstall, J. B. Prez, M.Caorsi, A. Medina-Mardones, A. Dassatti, K. Hess, Giotto-tda: A Topological Data Analysis Toolkit for Machine Learning and Data Exploration, arXiv. 2004.0255 (2020).

[9] V. de Silva, P. Skraba and M. Vejdemo-Johansson, Topological Analysis of Recurrent Systems, Workshop on Algebraic Topology and Machine Learning, NIPS 2012, Preprint available at http://sites.google.com/site/nips2012topology/contributed-talks

[10] F. Takens. Detecting strange attractors in turbulence. In: Dynamical Systems and Turbulence, Warwick 1980. Ed. by David Rand and Lai-Sang Young. Vol. 898. Lecture Notes in Mathematics, Springer Berlin/Heidelberg. (1981) 366–381.

[11] Kennel, Matthew B.; Abarbanel, Henry D. I. False neighbors and false strands: A reliable minimum embedding dimension algorithm. Physical Review E. 66(2) (2002) 026209.

[12] C. Shannon, The mathematical theory of communication. University of Illinois Press, Urbana, Illinois (1949).

[13] M.R. Muldoon, R.S. MacKay, J.P. Hu.keb and D.S. Broomhead, Topology from time series, Physica D. 65 (1993) 1–16.

[14] Perea, J.A., Harer, J. Sliding Windows and Persistence: An Application of Topological Methods to Signal Analysis. Found Comput Math. 15 (2015) 799–838.

[15] C. M. M. Pereira, F. M, Rodrigo, Persistent homology for time series and spatial data clustering, Expert Systems with Applications. 42(1516) (2015) 6026–6038.
[16] Fung, K.F., Huang, Y.F. & Koo, C.H. Assessing drought conditions through temporal pattern, spatial characteristic and operational accuracy indicated by SPI and SPEI: case analysis for Peninsular Malaysia. Nat Hazards. 103 (2020) 2071–2101.
[17] S. Aghabozorgi, A. Seyed Shirkhorshidi, Y.W. Teh,Time-series clustering a decade review. Information Systems. 53 (2015) 16–38.
[18] Everitt, B.S., Dunn, G., 2001. Applied Multivariate Data Analysis. 2^{nd} ed., John Wiley & Sons.
[19] Everitt, B.S., Landau, S., Leese, M., Stahl, D., 2011. Cluster Analysis. 5^{th} ed., John Wiley & Sons.
[20] Johnpaul, C., Prasad, M. V., Nickolas, S., & Gangadharan, G. Trendlets: A novel probabilistic representational structures for clustering the time series data. Expert Systems with Applications. 145 (2020) Article 113119.
[21] Paparrizos, J., & Gravano, L. Fast and accurate time-series clustering. ACM Transactions on Database Systems. 42(2) (2017) 8:18:49.
[22] Rakthanmanon, T., Campana, B., Mueen, A., Batista, G., Westover, B., Zhu, Q.,Zakaria, J., & Keogh, E. Searching and mining trillions of time series subsequences under dynamic time warping. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discover (2012).
[23] de Gois, G., de Oliveira-Jnior, J.F., da Silva Junior, C.A. et al. Statistical normality and homogeneity of a 71-year rainfall dataset for the state of Rio de JaneiroBrazil. Theor Appl Climatol. 141 (2020) 15731591.
[24] Herbert Edelsbrunner and John Harer. Computational Topology: an Introduction. AMS Press, Durham, North Carolina, 2009.