Malaysian Journal of Science 38 (Special Issue 2): 14- 22 (2019)

**THE INTERNATIONAL SEMINAR ON MATHEMATICS IN INDUSTRY (ISMI)
AND THE INTERNATIONAL CONFERENCE ON THEORETICAL AND APPLIED STATISTICS (ICTAS)
ISMI-ICTAS18 [4-6 SEPTEMBER 2018]**

Assessing Dynamic-Time-Warping Dissimilarity Measures in Regionalization of River Discharges

Nur Syazwin Mansor[1a], Norhaiza Ahmad[1b]*, Arien Heryansyah[2c]

[1] Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, 81310 Johor Bahru, MALAYSIA. E-mail: nsyazwin3@live.utm.my[a] ; norhaiza@utm.my[b]
[2] Faculty of Engineering, Universitas Ibn Khaldun (UIKA), Bogor. Jalan KH Sholeh Iskandar KM.2, Kedung Badak, Tanah Sereal, Kota Bogor, Jawa Barat 16162, INDONESIA. Email: arengga@gmail.com[c]
* Corresponding Author: norhaiza@utm.my[b]
Received: 21st April 2019 Revised : 6th August 2019 Published: 30th September 2019
DOI : https://doi.org/10.22452/mjs.sp2019no2.2

ABSTRACT Regionalization of river discharges is a process of transferring hydrological information to generalize hydrological information from one river to another. One approach to regionalize river discharge is to use a distance-based regional analysis by employing a Dynamic Time Warping (DTW) dissimilarity measure to cluster homogeneous river discharge patterns based on sequenced of time series discharge data. However, clustering homogeneous river discharge patterns can be sensitive to the choice of distance metric measures used due to out of phase behavior in the discharge time series. In this study, we assess three types of Dynamic Time Warping (DTW) measures specifically conventional DTW, a feature based DTW and a weighted based DTW on four annual discharge time series from six rivers in the state of Johor, Malaysia. A comparison of eight different clustering validation indices to determine the optimal number of rivers clusters with similar discharge patterns. These indices are used to measure the internal and external strength of the identified clusters. The results indicate that weighted based DTW outperform the conventional DTW and feature based DTW with 75% of the clustering indices agree that there are three optimal clusters of river discharge. By using weight as a function in DTW, it helps to cater the out of phase behavior in river discharge time series with the highest agreement of clustering indices compared to other types of DTW measures. We also found that three of the rivers (Sayong, Bekok, and Segamat) have similar river discharge patterns and could be used together in the generalization process. Meanwhile, the other rivers (Johor, Kahang, and Muar) varies in their time series patterns.

Keywords: Dynamic Time Warping, Clustering, Dissimilarity measure

1. INTRODUCTION

Regionalization refers to a process of transferring hydrological information to generalize hydrological information from one river to another (Razavi & Coulibaly, 2012; Snelder et al., 2005). It is essential not only to ensure the transferability of information when applying regionalization methods, but can also provide valuable indications to improve the understanding of the dominant physical phenomena in the different groups (Toth, 2013). For instance, similar river discharge can be classified according to their annual maximum flows, coefficient of variation and skewness of annual maximum discharges, latitude and longitude of selected stations (Agarwal et al., 2016; Corduas, 2011; Dikbas et al., 2013; Elesbon et al., 2015). Such information can be used to estimate low discharge magnitude and frequency in river streams for the purpose of river management

(Kahya et al., 2007; Laaha & Bloschl, 2006).

One approach of regionalization is through a time-based cluster analysis. It is used to cluster homogeneous river discharge patterns based on sequence of time series discharge data. This approach classifies a set of individual time series which possess similar patterns, shapes, or changes through selected interval period of time (Ouyang et al., 2010). Each data points of the time series are objects of ordered time sequences. This time-based clustering takes into account of the whole sequence of river discharge time series and is analyzed to identify the similarity across different sets of time series discharge data.

In time-based clustering, a nonlinear dissimilarity measure is commonly considered. A common nonlinear mapping dissimilarity measure used by hydrologists for ordered time sequence data is Dynamic Time Warping (DTW) dissimilarity measure (Gertseema et al., 2016; Gupta & Chaturvedi, 2013; Mishra et al., 2015). This nonlinear dissimilarity measure aligns each data points elastically which allows similar shapes to match when the points are out of phase in a time series sequence. It does not only take into account the smallest distance between data points of each discharge sequences but also the phase-difference of the two sequences. This method allows similar shape to be matched and the variability of the river discharge is considered by warping the time of discharge occurrences.

In the literature, there are several variations of DTW. However, as far as we are concerned only the conventional DTW has been applied to analyze the river discharge data to date. Therefore, in this study, we assess three types of Dynamic Time Warping (DTW) measures specifically conventional DTW, a feature based DTW and a weighted based DTW in the regionalization of river discharge data.

2. RIVER DISCHARGE DATA

The river discharge data used in the analysis were obtained from the Malaysian Drainage and Irrigation Department which consist of the daily observed of six rivers (Johor, Sayong, Bekok, Kahang, Muar, and Segamat) over 34 years from 1980 to 2014. However, after the missing records are removed, only 28 years remain in the dataset from 1980 until 2008. In this study, only four annual discharge data (1981, 1982, 1983, and 1987) for each river were used.

River discharge time series of each river for 1981 are plotted in Figure 1. At a glance, it can be seen that there are several homogeneous groups of river. First, based on the river discharge pattern, it can be observed that Sungai Johor and Sungai Kahang share the same pattern of river discharge which consistently fluctuate throughout the year and have a very high peak at the end of the year. Sungai Sayong and Sungai Bekok appear to share the same pattern of river discharge where there is a small peak during the month of May-June and moderately high in December. However, this assumption cannot be verified only by visualization. It needs to be statistically proven. The process of identifying homogenous river discharge pattern using a time-based cluster analysis is described in the following section.

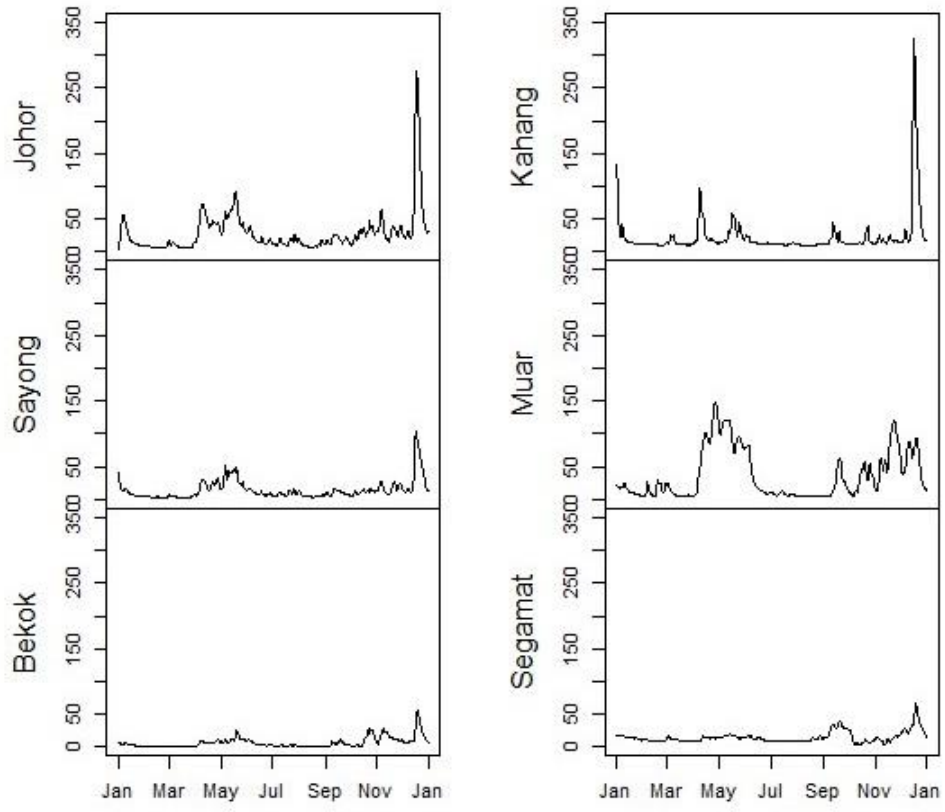

Figure 1: River discharge of six rivers in 1981.

3. METHODOLOGY

In general, the process of identifying the temporal grouping of river discharge process using cluster analysis comprises of four major steps: transformation of raw data, dissimilarity measure between annual river discharge, clustering algorithm to identify the membership of cluster, and validation index to validate the cluster membership.

In this study, transformation is not required since the measurement scale of river discharge are the same for each year. The homogeneous annual discharge processes are then identified by comparing three different types of Dynamic Time Warping (DTW) measures specifically conventional DTW, a feature based DTW and a weighted based DTW. Identification of regional grouping of river discharge is done using K-medoid clustering algorithm and C-Index measure to validate the membership of clusters.

3.1 Dynamic Time Warping (DTW)

Dynamic Time Warping (DTW) is an algorithm for measuring optimal similarity between two river discharge time sequences. The time series data vary not only on the time amplitudes but also in terms of time progression (Mishra et al., 2015). The non-linear alignment produces a similar measure, allowing similar shapes to match even if they are out of phase in the time axis. Sequences are "warped" non-linearly in the time dimension to determine their measure of similarity independence for certain non-linear variations in the time dimension. In respect of river discharge time series data, the DTW algorithm for dissimilarity measure is as follow:

$$Y_{1j} = y_{11}, y_{12}, ..., y_{1m} \quad (1)$$

$$Y_{2j} = y_{21}, y_{22}, ..., y_{2m} \quad (2)$$

An $m \times n$ matrix is constructed using DTW aligning for these two sequences. Each element in the matrix contains the dissimilarity between two points Y_{1j} and Y_{2j}, called Euclidean distance, D_{EUC}. The distance is defined as:

$$D_{EUC}(Y_i, Y_j) = \sqrt{\sum_{k=1}^{p} \left(Y_{ik} - Y_{jk}\right)^2} \quad (3)$$

where Y_{ik} and Y_{jk} are respectively the k^{th} river discharge value of the p-dimensional monsoon period for individuals i and j. A warping path, W is a contiguous set of matrix elements that defines a mapping between Y_{1j} and Y_{2j}. The kth element of W is defined as $w_k = (i, j)_k$, so we have:

$$W = w_1, w_2, ..., w_k, ..., w_K \text{ where } (m, n) \le K < (m + n - 1) \quad (4)$$

The warping path is typically subject to several constraints:

i. Boundary conditions: $w_1 = (1,1)$ and $w_k = (m, n)$. The warping path needs to start and finish diagonally opposite corner cells of the matrix.
ii. Continuity: Given $w_k = (a, b)$ then $w_{k-1} = (a', b')$, where $a - a' \le 1$ and $b - b' \ge 1$. The allowable steps in the warping path are restricted to adjacent cells (including diagonally adjacent cells).
iii. Monotonicity: Given $w_k = (a, b)$ then $w_{k-1} = (a', b')$, where $a - a' \ge 0$ and $b - b' \le 0$. The points in W need to be monotonically spaced in time.

Many warping paths satisfy the constraints, but only one path is chosen which minimizes the warping cost taken by:

$$D_{DTW}(X_1, X_2) = \min\left(\frac{1}{k}\sum_{k=1} w_k\right) \quad (5)$$

where k in the denominator used to compensate the fact that warping paths may have different lengths.

3.2 Derivative Dynamic Time Warping (DDTW)

Derivative Dynamic Time Warping (DDTW) utilize information on the shape of the time series by considering the first derivative of the sequences. Previously, the conventional DTW construct $m \times n$ path matrix where the matrix contains the dissimilarity between two points Y_{1j} and Y_{2j}, called Euclidean distance, D_{EUC}. With DDTW the dissimilarity measure is not Euclidean but rather the square of the estimated derivatives of Y_{1j} and Y_{2j}. Estimate derivatives of each data points are:

$$D_{DER}[Y] = \frac{(y_i - y_{i-1}) + ((y_{i+1} - y_{i-1})/2)}{2} \quad (6)$$

This estimate is simply the average of the slope of the line through the point in question and its left neighbor, and the slope of the line through the left neighbor and the right neighbor. Empirically this estimate is more robust to outliers than any estimate considering only two data sequences. Note the estimate is not defined for the first and last elements of the sequence. Instead, we use the estimates of the second and penultimate elements respectively.

3.3 Weighted Dynamic Time Warping (WDTW)

Weighted Dynamic Time Warping (WDTW) is the modified version of the conventional DTW as described in 3.1 which calculates the distance of all pairwise points with equal weight of each point regardless of phase difference, WDTW penalizes the points according to the phase difference between discharge of the two river discharge time series (Jeong et al., 2011). In WDTW algorithm, when creating an $m \times n$ path matrix, the distance between the two points y_i and y_j is calculated as:

$$D_W(y_i, y_j) = \left\| w_{|i-j|}(y_i - y_j) \right\| \quad (7)$$

where $w_{|i-j|}$ is the positive weight value between the two points y_i and y_j. A Modified Logistic Weight Function (MLWF) is used to assign weight and the weight value $w_{(i)}$ is defined as:

$$w_{(i)} = \left[\frac{w_{\max}}{1 + \exp(-g(i - m_c))} \right] \quad (8)$$

where $i = 1, \ldots, m$, m is the length of a sequence and m is the midpoint of a sequence. $w_{\max}$ is the desired upper bound for the weight parameter, and g is an empirical constant that controls the curvature (slope) of the function; that is, g controls the level of penalization for the points with larger phase difference. The value of $w_{\max}$ is set to 1 and g is set to 0.6 as recommended by Jeong et al. (2011).

3.4 K-medoid Clustering

Instead of using the mean point as the center of a cluster, K-medoids uses an actual point in the cluster to represent it. Medoid is the most centrally located object of the cluster, with minimum sum of distances to other points. First, we select K representative points to form initial clusters, and they are repeatedly moved to better cluster representatives. All possible combinations of representative and non-representative points are analyzed, and the quality of the resulting clustering is calculated for each pair. An original representative point is replaced with the new point which causes the greatest reduction in distortion function. At each iteration, the set of best points for each cluster form the new respective medoids.

3.5 Internal Validity Measure

Internal validity measures are evaluated based on the compactness and separation. Compactness refers on how close the annual

river discharges within a cluster. Meanwhile, separation refers to how distinct a cluster is from other clusters of river discharges. In this study, we use eight different internal validity measures which are Calinski-Harabasz (CH), Silhouette, Dunn, C, Davies-Bouldin (DB), McClain, Krzanowski-Lai (KL), and S_Dbw indices (Desgraupes, 2013; Zhao, 2012).

4. RESULTS AND DISCUSSION

We use four years of annual discharge for each six rivers in Johor as described in Section 2. This dataset is reflected by a 365×24 data matrix from the corresponding vectors. Three dissimilarity measures (DTW, DDTW, and WDTW) are applied to the dataset to produce different 24×24 dissimilarity matrices. Each dissimilarity matrix is used in the K-medoid clustering procedure to extract different cluster solution from the river discharge. In order to find the optimal number of clusters, we tested different number of cluster solutions ranging from three to five clusters based on its internal clustering indices.

Table 1 shows the internal indices value for three to five cluster solution. The first three indices (CH, Silhouette, and Dunn) indicate strong similar river discharge pattern when the values are large. The larger the value, the more accurate closeness within each cluster. In contrast, the next five indices (C, DB, McClain, KL, and SDBW) indicate strong similar river discharge pattern when the values are small. The smaller the value, the more accurate closeness within each cluster.

We summarize the agreement between all the indices in the last row of Table 1. WDTW clearly outperforms the other two dissimilarity measures with 75% of the clustering indices agree that there are three optimal cluster solutions for the river discharge. Meanwhile, only 25% of the indices agree that there are three or five optimal numbers of clusters using DTW and DDTW respectively.

Table 1: Internal clustering indices for 3 to 5 cluster solution.

Index/No. of Clusters	**3**			**4**			**5**		
	DTW	DDTW	WDTW	DTW	DDTW	WDTW	DTW	DDTW	WDTW
CH	6.51	6.16	**6.97**	5.64	5.27	5.96	5.79	5.49	5.44
Silhouette	0.29	0.29	**0.47**	0.30	0.30	0.36	0.30	0.30	0.36
Dunn	0.29	0.29	**0.45**	0.37	0.36	0.33	0.43	0.43	0.41
C	**0.29**	0.30	0.43	0.34	0.34	0.39	0.48	0.49	0.42
DB	1.63	1.67	**0.99**	1.63	1.68	1.62	1.46	1.51	1.60
McClain	0.94	0.96	**0.26**	0.99	1.01	0.93	1.02	1.04	1.09
KL	2.37	2.46	2.09	0.82	**0.75**	1.18	2.15	2.21	1.26
SDBW	0.98	1.01	**0.49**	0.93	0.94	0.53	0.66	0.67	0.65
Percentage of agreement (%)	12.5%	-	75%	-	12.5%	-	-	-	-

In the following Table 2, we detailed out the membership of the three cluster solution using WDTW dissimilarity measure. Cluster 1 consist of 11 annual river discharge, Cluster 2 consist of 8 annual river discharge, and Cluster 3 consist of only 5 annual river discharge. In particular, Sayong, Bekok, and Segamat river are categorized in Cluster 1. These three rivers are located at the North of Johor. This implies that they have similar river discharge patterns and could be used together in the generalization process. Meanwhile, the other rivers (Johor, Kahang, and Muar) in the other clusters vary in their time series patterns.

Table 2: Membership of clusters using WDTW dissimilarity measure.

Rivers	**Cluster 1**	**Cluster 2**	**Cluster 3**
Johor river discharge 1981		X	
Johor river discharge 1982			X
Johor river discharge 1985			X
Johor river discharge 1987			X
Sayong river discharge 1981	**X**		
Sayong river discharge 1982		X	
Sayong river discharge 1985	**X**		
Sayong river discharge 1987	**X**		
Bekok river discharge 1981	**X**		
Bekok river discharge 1982	**X**		
Bekok river discharge 1985	**X**		
Bekok river discharge 1987	**X**		
Kahang river discharge 1981		X	
Kahang river discharge 1982		X	
Kahang river discharge 1985	**X**		
Kahang river discharge 1987		X	
Muar river discharge 1981		X	
Muar river discharge 1982		X	
Muar river discharge 1985			X
Muar river discharge 1987			X
Segamat river discharge 1981	**X**		
Segamat river discharge 1982		X	
Segamat river discharge 1985	**X**		
Segamat river discharge 1987	**X**		
Total	**11**	8	5

5. CONCLUSION

In this study, we assess three different types of Dynamic Time Warping (DTW) dissimilarity measures in regionalization of river discharge data. Weighted DTW (WDTW) was found to be superior compared to the other two types of DTW (conventional DTW and feature based DTW). Although the six rivers are located in the same state, they have different fluctuation pattern throughout the year. This may result to unaligned mapping of river discharge time series due to slightly higher/lower feature between sequences when similarities are sought. Hence, it could lead to singularities problem where a single point on one-time series maps onto a large subsection of another time series (Jeong et al., 2011).

Thus, weighted DTW helps to re-align the higher or lower feature between sequences

by penalizing the points according to the phase difference between river discharge time series. This caters the issue of small time shift between discharge peaks in the main river and in its tributaries as highlighted by Gertseema et al. (2016). However, the types of DTW dissimilarity measure compared in this study used raw river discharge data or only local feature of discharge data that represent relationship with two adjacent neighboring points. Further research will be conducted on studying a better feature to be included in DTW dissimilarity measure that is able to include the overall significant or global features that occur in the sequence of river discharge data.

6. REFERENCES

Agarwal, A.; Maheswaran, R.; Sehgal, V.; Khosa, R.; Sivakumar, B. & Bernhofer, C. (2016). Hydrologic regionalization using wavelet-based multiscale entropy method. *Journal of Hydrology,* 538: 22-32.

Corduas, M. (2011). Clustering streamflow time series for regional classification. *Journal of Hydrology,* 407: 73-80.

Desgraupes, B. (2013). Clustering indices. *University of Paris Ouest-Lab Modal'X,* 1: 34.

Dikbas, F.; Firat, M.; Koc, A. C. and Gungor, M. (2013). Defining homogeneous regions for streamflow processes in Turkey using a K-means clustering method. *Arabian Journal for Science and Engineering,* 38: 1313-1319.

Elesbon, A. A.; Silva, D. D. d.; Sediyama, G. C.; Guedes, H. A.; Ribeiro, C. A. and Ribeiro, C. B. d. M. (2015). Multivariate statistical analysis to support the minimum streamflow regionalization. *Engenharia Agricola, SciELO Brasil,* 35: 838-851.

Geertsema, T.; Teuling, A.; Uijlenhoet, R.; Torfs, P.; Hoitink, A. and Weerts, A. (2016). Simultaneous occurrence of discharge peaks in a large river and its lowland tributaries. In *River Flow - Proceedings of the International Conference on Fluvial Hydraulics,* St. Louis, 11-14 July 2017, pp. 1626-1630.

Gupta, A. and Chaturvedi, S. K. (2013). Real Time Prediction System of Discharge of the Rivers using Clustering Technique of Data Mining. *International Journal of Engineering Research and Development,* 9: 12-24.

Jeong, Y. S., Jeong, M. K., and Omitaomu, O. A. (2011). Weighted dynamic time warping for time series classification. *Pattern Recognition,* 44(9), 2231-2240.

Kahya, E.; Demirel, M. Cü. and Piechota, T. C. (2007). Spatial grouping of annual streamflow patterns in Turkey. *27th AGU Hydrology Days,* 169-176

Laaha, G. and Blöschl, Gü. (2006). A comparison of low flow regionalisation methods: catchment grouping. *Journal of Hydrology, Elsevier,*323: 193-214.

Mishra, S.; Saravanan, C.; Dwivedi, V. and Pathak, K. (2015). Discovering flood rising pattern in hydrological time series data mining during the pre monsoon period. *Indian Journal of Geo-Marine Sciences,* 44: 3.

Ouyang, R.; Ren, L.; Cheng, W. and Zhou, C. (2010). Similarity search and pattern discovery in hydrological time series data mining. *Hydrological Processes,* 24(9): 1198-1210.

Razavi, T. & Coulibaly, P. (2012). Streamflow

prediction in ungauged basins: review of regionalization methods. *Journal of Hydrologic Engineering, American Society of Civil Engineers,*18: 958-975.

Snelder, T. H.; Biggs, B. J. & Woods, R. A. (2005). Improved eco-hydrological classification of rivers. *River Research and Applications*, 21(6): 609-628.

Toth, E. (2013). Catchment classification based on characterisation of streamflow and precipitation time series. *Hydrology and Earth System Sciences*, 17: 1149-1159.

Zhao, Q. (2012). *Cluster Validity in Clustering Methods*. Dissertations in Forestry and Natural Sciences, University of Eastern Finland.