MJFAS MALAYSIAN JOURNAL OF FUNDAMENTAL AND APPLIED SCIENCES
PRINT ISSN: 2289-5981 | ONLINE ISSN: 2289-599X

RESEARCH ARTICLE

Spatial grouping of homogeneous river flow process in Johor

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

[a] Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, 81310 Skudai, Johor Darul Takzim, Malaysia.
[b] Centre or Climate Risk and Opportunity Management in Southeast Asia Pasific (CCROM SEAP), Gedung Fisik dan Botani, Lantai 2 Kampus IPB Baranangsiang, Jalan Pajajaran Bogor, 16143 Jawa Barat Indonesia.

* Corresponding author: nsyazwin3@live.utm.my, norhaiza@utm.my

Article history
Received
Accepted 8 November 2017

Abstract

This study identifies the spatial grouping of homogeneous river flow process on eight rivers in Johor based on river discharge records of 28 years. A clustering approach using a nonlinear dissimilarity measure called Dynamic Time Warping (DTW) is used to detect the similarity between these eight rivers. The clustering results validated by an internal validity measure, C-index shows two distinct groups of rivers: Cluster 1 consist of Sungai Lenik, Sungai Segamat, Sungai Bekok, and Sungai Muar; Cluster 2 consist of Sungai Sayong, Sungai Lenggor, Sungai Johor, and Sungai Kahang. This two-cluster solution are stable and interpretable with reference to spatial variations and can be distinguish by their geographical location in the peninsular.

Keywords: Spatial grouping, Clustering, Dissimilarity measure

INTRODUCTION

River flow information are important in defining availability of water supply, evaluation of flood control, used in irrigation process, source of hydropower plant, etc. In Malaysia, reliable river flow information can be compromised by incomplete information due to the existence of many ungauged catchments (Razaqa *et al*., 2016). Thus, identification of group of hydrologically similar catchment is useful to ensure transferability of information when applying region classification. In hydrology, such processes is also known as regionalization methods which resulting groups of catchments with similar or homogenous river flow. This method provide valuable indications to improve the understanding of the dominant physical phenomena in the different groups (Sawicz *et al*., 2011).

There are numerous studies identifying spatial or region with homogeneous river flow pattern specifically using cluster analysis. For instance, Mediero *et al*. (2015) defined five homogeneous river regions by using a hierarchical clustering algorithm and each homogeneous cluster is proportional to its physiographic characteristics, climate, and land-use pattern. Clustering method also been used by Dikbas *et al*. (2013) and Kahya *et al*. (2008) for region classification according to annual maximum flows, coefficient of variation and skewness of annual maximum flows, latitude and longitude of selected stations. Many studies including the above would use linear dissimilarity measure such as euclidean to identify hydrologically similar river flow pattern. Although this method has been shown to detect varied groups of rivers, however such method might not suitable when analyzing river discharge records. This is due to the nonlinearity nature of river flow process (Wang *et al*., 2016) and the existence of temporal dependency between observations.

Therefore, this study aims to identify the spatial grouping of homogeneous river flow process on eight rivers in Johor by using cluster analysis. First, we transform the river discharge time series data to frequency domain using Discrete Fourier Transform (DFT). Second, we incorporate a nonlinear dissimilarity measure, Dynamic Time Warping (DTW) to quantify closeness of the transformed river discharge data. The resulting dissimilarity of river discharge is then used for K-means clustering.

MATERIALS AND METHODS

River flow data

In this study, the river flow data represented by daily discharge data of the eight rivers in Johor were obtained from the Department of Irrigation & Drainage, Malaysia from 1980 until 2013. Rivers used in this study are Sungai Muar, Sungai Segamat, Sungai Lenik, Sungai Bekok, Sungai Kahang, Sungai Lenggor, Sungai Sayong, and Sungai Johor as illustrated in Figure 1. However, only complete dataset of 1980 until 2008 are taken.

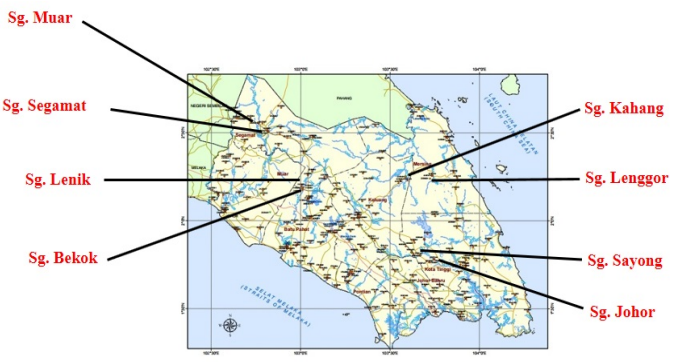

Figure 1 Eight rivers' station in Johor (Department of Irrigation & Drainage).

River discharge time series of each rivers are plotted in Figure 2. At a glance, it can be seen that there are several homogeneous groups of river. First, based on the river flow pattern, it can be observed that Sungai Johor, Sungai Sayong and Sungai Muar share the same pattern of river flow which consistently fluctuate over time. Meanwhile, Sungai Lenggor and Sungai Segamat share the same low flow distribution and only peak at certain period of time. However, this assumption cannot be verified only by visualization but need to be statistically proven. The processes of identifying homogenous river flow pattern using cluster analysis are described in the following section.

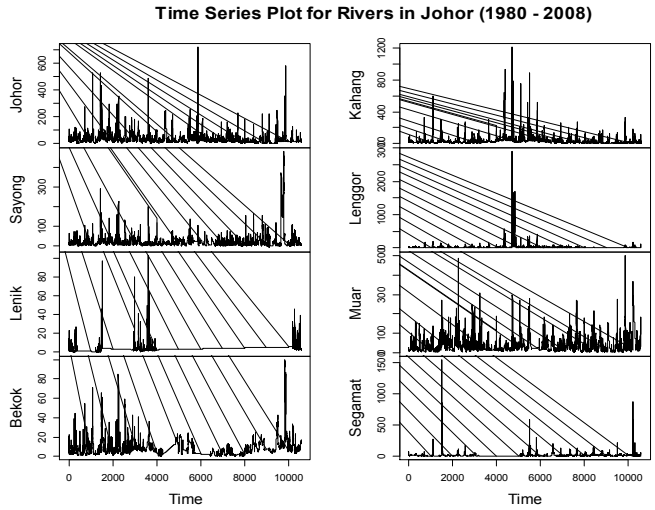

Figure 2 Time series plot of eight rivers in Johor.

Clustering of homogeneous river flow

To cluster the rivers in Johor into its homogeneous river flow pattern, some consideration on the choice of clustering component need to be done. In general, clustering approach has four components: data pre-process, dissimilarity measure, clustering algorithm and cluster validation as shown in Figure 3.

Figure 3 Clustering component

Discrete Fourier Transform (DFT) is used in this study to transform the discharge data from time domain into frequency domain. The transformation of discharge data reduces the dimensionality of the original data. Each frequency X_F is calculated as:

$$X_F = \sum_{n=0}^{N-1} x_n e^{-\frac{i2\pi fn}{N}} \quad (1)$$

where
N = number of river discharge time series
n = current time series we are considering (0,...,N-1)
x_n= value of the river discharge at time n
f = current frequency we are considering (0H_z up to N-1H_z)
X_F = amount of frequency in the signal

Dynamic Time Warping (DTW) is a nonlinear algorithm for measuring optimal similarity between two river discharge time sequences [6]. Let R_1 and R_2 be the two time series river discharge sequences of length m and n respectively, given as:

$$R_1 = x_1, x_2, \dots, x_m \quad (2)$$
$$R_2 = y_1, y_2, \dots, y_n \quad (3)$$

An $m \times n$ matrix is constructed using DTW, aligning these two sequences, R_1 and R_2. Each element in the matrix contains the distance between two points x_i and y_j, called Euclidean distance. A warping path, W, is a contiguous set of matrix elements that defines a mapping between R_1 and R_2. The kth element of W is defined as $w_k = (i,j)_k$, so we have:

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

The warping path is typically subject to several constraints:

i. Boundary conditions: $w_1 = (1,1)$ and $w_k = (m,n)$. The warping path need to start and finish in 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 is 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:

$$D_{DTW}(R_1, R_2) = \min\left(\frac{1}{k}\sum_{k=1} w_k\right) \quad (5)$$

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

K-means clustering method is used to identify spatial groupings of the rivers. First, the centers of K number of clusters are determined and each variable is assigned to the nearest cluster center with the help of a dissimilarity measure. After the assignment of each variable in the input data set to a cluster, the cluster centers for all clusters are recalculated and the variables might be assigned to different clusters according to the locations of the new cluster centers. This process is repeated until there is no change in cluster centers.

C-index measure is used to measure the internal validity between the clusters for validation purposes. The index validation is defined as:

$$C = \frac{S - S_{min}}{S_{max} - S_{min}} \quad (6)$$

where S is the sum of distances over all pairs of objects form the same cluster, n is the number of those pairs and S_{min} is the sum of the n smallest distances if all pairs of objects are considered. Likewise S_{max} is the sum of the n largest distances out of all pairs. The C-index is limited to the interval [0, 1] and should be minimized. Figure 4 shows the summary or flowchart of each clustering process used in this study.

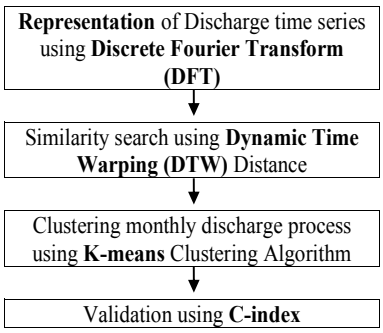

Figure 4 Flowchart of clustering process

RESULTS AND DISCUSSION

The identification of spatial grouping of homogeneous river flow process on eight rivers in Johor utilize K-means clustering technique which incorporate a nonlinear dissimilarity measure, Dynamic Time Warping (DTW). Initially, two to five cluster solutions are defined in K-means clustering algorithm. Then, we compared two to five cluster solution based on C-index measure with the limit guideline of 0 to 1. The smallest value of C-index indicates a good clustering which reflect the closeness within the membership in the cluster.

Table 1 shows that two cluster solution shows the lowest C-index value 0.405. Cluster 1 consist of Sungai Lenik, Sungai Segamat, Sungai Bekok, and Sungai Muar. Cluster 2 consist of Sungai Sayong, Sungai Lenggor, Sungai Johor, and Sungai Kahang as shown in Figure 5.

Table 1 C-index value of two to five cluster solution

No of Clusters	C-Index
2	0.405
3	0.549
4	0.687
5	0.745

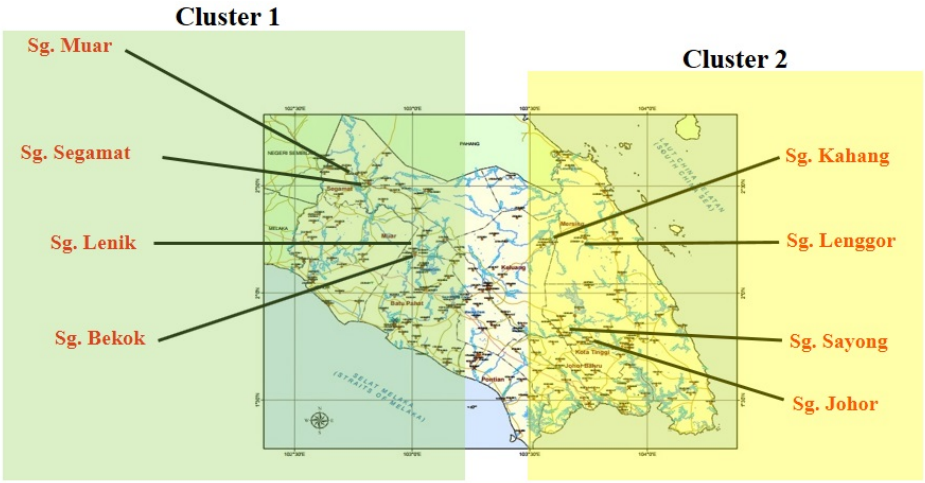

Figure 5 River location based on two cluster solution.

Figure 5 above illustrate the geographical distribution of the resultant two clusters. The green region is the cluster covering western coastal zone while the yellow region covering eastern coastal zone in Johor. The differences in river flow patterns across the two cluster regions might be explained through rainfall distribution during the southwest monsoon and northeast monsoon period. The two-cluster solutions are stable and interpretable regarding spatial variations.

CONCLUSION

The results of this study provided insights into the spatial grouping of homogeneous river flow patterns in Johor. The utilization of k-means clustering methodology has shown distinct cluster of rivers. It is recommended for future similar grouping studies to corporate a better dissimilarity measure which can capture the river flow characteristics in specific such as the peak of discharge pattern in the series.

ACKNOWLEDGEMENT

Ministry of High Education (MOHE), STEM Grant with vote no. A. J091002.5600.07397.

REFERENCES

Dikbas, F., Firat, M., Koc, A. C., & Gungor, M. (2013). Defining homogeneous regions for streamflow processes in Turkey using a K-means clustering method. Arabian Journal for Science and Engineering, 38(6), 1313-1319.

Isik, S., & Singh, V. P. (2008). Hydrologic regionalization of watersheds in Turkey. Journal of Hydrologic Engineering, 13(9), 824-834.

Kahya, E., Kalaycı, S., & Piechota, T. C. (2008). Streamflow regionalization: case study of Turkey. Journal of Hydrologic Engineering, 13(4), 205-214.

Mediero, L., Kjeldsen, T. R., Macdonald, N., Kohnova, S., Merz, B., Vorogushyn, S., Wilson, D., Alburquerque, T., Blöschl, G., Bogdanowicz, E., Castellarin, A., Hall, J., Kobold, M., Kriauciuniene, J., Lang, M., Madsen, H., Onuşluel Gül, G., Perdigão, R.A.P., Roald, L.A., Salinas, J.L., Toumazis, A.D., Veijalainen, N., & Óðinn Þórarinsson. (2015). Identification of coherent flood regions across Europe by using the longest streamflow records. Journal of Hydrology, 528, 341-360.

Mishra, S., Saravanan, C., Dwivedi, V. K., & Pathak, K. K. (2015). Discovering flood rising pattern in hydrological time series data mining during the pre monsoon period. Computer Applications, 44(March), 35–44.

Razaqa, S. A., Ismailb, T., Heryansyahb, A., Lawanb, U. F., Alamgirb, M., & Pourb, S. H. (2016). Streamflow Prediction in Ungauged Catchments in the East Coast of Peninsular Malaysia Using Multivariate Statistical Techniques. Jurnal Teknologi, 78(6-12), 43-49.

Sawicz, K., Wagener, T., Sivapalan, M., Troch, P. A., & Carrillo, G. (2011). Catchment classification: empirical analysis of hydrologic similarity based on catchment function in the eastern USA. Hydrology and Earth System Sciences, 15(9), 2895.

Wang, W., Vrijling, J. K., Van Gelder, P. H., & Ma, J. (2006). Testing for nonlinearity of streamflow processes at different timescales. Journal of Hydrology, 322(1), 247-268.