International Journal of Engineering & Technology, 7 (4) (2018) 3228-3237

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i4.20339
Research paper

Identifying homogeneous rainfall catchments for non-stationary time series using tops is algorithm and bootstrap k-sample Anderson darling test

Zun Liang Chuan [1] *, Noriszura Ismail [2], Wan Nur Syahidah Wan Yusoff [1], Soo-Fen Fam [3], Mohd Akramin Mohd Romlay [4]

[1] Faculty Industrial Sciences & Technology, University Malaysia Pahang, Lebuhraya Tun Razak, 26300 Gambang, Kuantan, Pahang Darul Makmur
[2] School of Mathematical Sciences, Faculty Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor Darul Ehsan
[3] Faculty of Technology Management and Technopreneurship, University Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100, Durian Tunggal, Melaka
[4] Faculty of Mechanical Engineering, University Malaysia Pahang, 26600 Pekan, Pahang Darul Makmur
**Corresponding author E-mail: chuanzunliang@ump.edu.my*

Abstract

The reliability of extreme estimates of hydro-meteorological events such as extreme rainfalls may be questionable due to limited historical rainfall records. The problem of limited rainfall records, however, can be overcome by extrapolating information from gauged to ungauged rainfall catchments, which requires information on the homogeneity among rainfall catchments. The purpose of this study is to introduce a new regionalization algorithm to identify the most suitable agglomerative hierarchical clustering (AHC) algorithm and the optimum number of homogeneous rainfall catchments for non-stationary rainfall time series. The new algorithm is based on the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) algorithm. This study also suggests the use of Bootstrap K-sample Anderson Darling (BKAD) test for validating regionalized homogeneous rainfall catchments. The Cophenetic Correlation Coefficients (CCC) from ten similarity measures are used as attributes for the TOPSIS algorithm to identify the most suitable AHC algorithm out of seven algorithms considered. The C-index (δ_{CI}), Davies-Bouldin index (δ_{DB}), Dunn index (δ_{DI}) and Gamma index (δ_{GI}) are then used as attributes for the TOPSIS algorithm to determine the optimum number of homogeneous rainfall catchments. The results show that the most suitable AHC algorithm is able to cluster twenty rainfall catchments in Kuantan River Basin, Malaysia into two optimum significant homogeneous clusters. The results also imply that the BKAD test is invariant towards the number of Bootstrap samples in the validation of homogeneous rainfall catchments.

Keywords: *Agglomerative Hierarchical Clustering Algorithm; Bootstrap K-Sample Anderson-Darling Test; Non-Stationary Time Series; TOPSIS Algorithm*

1. Introduction

The reliability of water resource management planning, hydraulic structure design and flood plain zoning are highly dependent on the extreme estimates of hydro-meteorological events such as extreme rainfalls. This reliability, however, is often restricted due to limited records of historical rainfalls of ungauged catchments. Several regionalization algorithms have been suggested to overcome this restriction by extrapolating information from gauged to ungauged catchments, such as the agglomerative hierarchical clustering (AHC) algorithm [2], [6], [10], [15], [17], [32], [35], [44], the principal component algorithm [5], [42], the canonical correlation algorithm [7] and the neural network algorithm [16], [26].

The AHC algorithm is widely used in the area of hydrological regionalization [11]. For examples, Burn et al. [6] applied the simple linkage clustering algorithm and heterogeneity measures to regionalize and validate homogeneous catchments in the West-Central Canada. Guttman [15] applied the average linkage and Ward's minimum variance clustering algorithms to regionalize precipitation catchments in the United States, and validated the homogeneous catchments using the discordance and heterogeneity measures. Venkatesh and Jose [44] regionalized and validated homogeneous rainfall catchments in the Western Ghats using the Ward's minimum variance clustering algorithm and the analysis of variance. In the later years, Ngongondo et al. [32] and Pansera et al. [35] proposed an efficient two-stage clustering algorithm to regionalize homogeneous catchments of the Southern Malawi and Brazil respectively, where the homogeneous catchments were validated by the discordance and heterogeneity measures.

Several studies on regionalization of homogeneous catchments in Malaysia were also carried out. Ahmad et al. [2] concluded that the complete linkage clustering algorithm based on the correlation similarity metric is the most appropriate algorithm to regionalize homogeneous catchments in Peninsular Malaysia, and suggested that the optimum number of clusters to be determined when the majority of internal clustering validation indices show similar results. In another study, Hamdan et al. [17] regionalized rainfall patterns using the rainfall amount curves, and regionalized and validated homogeneous catchments using the complete linkage

clustering algorithm and the adaptive Neyman test. Recently, Chuan et al. [10] proposed another efficient regionalized algorithm in identifying homogeneous precipitation catchments in Kuantan River Basin, Pahang. Their proposed regionalized algorithm is the associated between the average linkage hierarchical clustering algorithm and multi-scale bootstrap resampling.

Most of the previous studies showed that the identification of the most appropriate AHC algorithm and the optimum number of homogeneous catchments is applied to stationary rainfall time series. On the other hand, the discordant and heterogeneity measures are more suitable for low skewed data [45]. Therefore, a new regionalization algorithm which is more suitable for non-stationary extreme rainfall time series is essential as past studies have shown that regional phenomenon, such as the monsoon, El Nino-Southern Oscillation, Indian Ocean Dipole and Madden-Julian Oscillation, have created non-stationary components in climate variability [1], [43].

The main objective of this study is to propose a new regionalization algorithm to identify the most appropriate AHC algorithm and the optimum number of homogeneous catchments for non-stationary rainfall time series using the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) algorithm. The use of Bootstraps K-sample Anderson Darling (BKAD) test is also proposed in this study to validate regionalized homogeneous catchments. The rest of this article is organized as follows: Section 2 provides the description of the study areas, while the methodologies for analyzing the monthly historical rainfall data are described in Section 3. Section 4 discusses the results. The concluding remarks and future works are presented in Section 5.

2. Study areas

Pahang is considered as one of the substantial districts of agricultural land use in Malaysia [31]. The state is often exposed to risks of flood occurrence during the Northeast Monsoon, which could result in massive impacts in terms of economic damages and fatalities. Even though Malaysia has transforms into a relatively open state-oriented and a newly industrialized market, the agriculture sector remains to play a significant role in ensuring food security, economic growth, socio-economic improvement, employment generation and poverty reduction of the nation [3], [14], [20]. Therefore, intelligent and adequate water resource management planning, hydraulic structure design and flood plain zoning are highly required to secure the agriculture activities which are supposedly unaffected by the quality and quantity of the water supply.

The monthly historical rainfall records from twenty rainfall catchments in Kuantan River Basin, Malaysia are considered in this study. The Kuantan River Basin is known as one of the significant tributaries that irrigates the majority of the rural, urban, agriculture and industrial areas of Kuantan District [49]. The locations of the twenty rainfall catchments are shown in Fig. 1, and the information on each station are illustrated in Table 1. The monthly rainfall records cover the period of February 2010 until November 2014, and are obtained from Department of Irrigation and Drainage (DID), Malaysia, which coverage period of the Northeast Monsoon. Despite the short period of monthly rainfall records is used in this study, however, this sample size is sufficient for risk assessment [25].

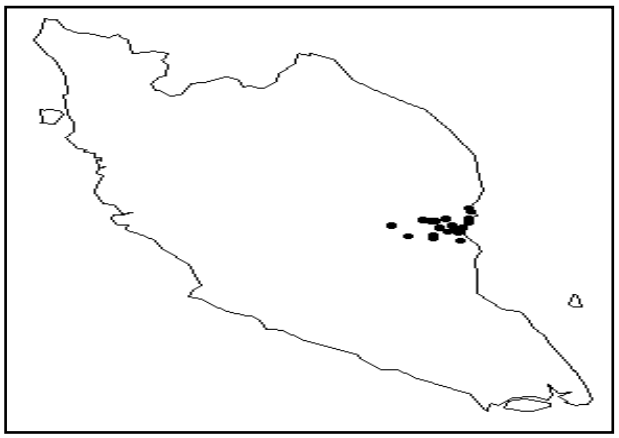

Fig. 1: Location of 20 Selected Rainfall Catchments in Kuantan River Basin, Pahang.

Table 1: Characteristics and Summary Statistics of Monthly Rainfall Historical Records of 20 Selected Rainfall Catchments in Kuantan River Basin, Pahang

Stn	Stn Name	Ev	Lat	Long	MN	CV
01	Sri Damai	14.9	03°44’47’’	103°13’20’’	90.5552	165.6909
02	Paya Bungor	34.7	03°41’30’’	102°56’00’’	158.1138	77.1298
03	Kampung Pulau Manis	37.4	03°39’10’’	103°07’10’’	181.5224	71.5749
04	Kampung Bahru	7.6	03°37’50’’	103°18’55’’	179.6224	92.8763
05	JKR Gambang	41.3	03°42’20’’	103°07’00’’	234.0069	66.1214
06	Paya Besar	6.0	03°46’20’’	103°16’50’’	162.8810	93.1208
07	Kampung Sungai Soi	11.9	03°43’50’’	103°18’00’’	210.9621	94.1730
08	Ladang Ulu Lepar	91.7	03°50’25’’	102°48’00’’	167.2017	65.7289
09	Ladang Mentiga	9.4	03°48’58’’	103°19’30’’	199.1931	73.1688
10	Panching	71.4	03°48’53’’	103°09’38’’	234.0707	86.9853
11	Paya Pinang	6.7	03°50’30’’	103°15’30’’	209.7328	93.7746
12	JPS Pahang	10.3	03°48’30’’	103°19’45’’	180.8603	97.8549
13	Ladang Jeram	-1.4	03°53’40’’	103°23’00’’	210.9414	119.5480
14	Sungai Lembing	33.1	03°55’00’’	103°02’10’’	245.6690	62.2638
15	Ladang Nada	16.9	03°54’30’’	103°06’20’’	227.9431	73.9733
16	Ladang Kuala Reman	29.9	03°54’00’’	103°08’00’’	201.8638	74.4579
17	Balok	4.1	03°56’40’’	103°23’00’’	220.8241	110.4599
18	Bukit Sagu	20.9	03°56’14’’	103°12’52’’	511.7517	79.6973
19	Kampung Cherating	9.0	04°05’35’’	103°22’50’’	221.2155	108.1321
20	Kampung Sungai Ular	58.5	04°30’00’’	103°23’40’’	228.7362	108.5308

*Note: Stn: Station; Ev: Elevation (In Meters); Lat: Latitude (North); Long: Longitude (East); MN: Monthly Average of Rainfall Amount (in Millimeters); CV: Coefficient of Variation (percentage).

3. Methodologies

An overview of the proposed algorithms to identify the most suitable AHC algorithm and the optimum number of homogeneous catchments is presented in Fig. 2.

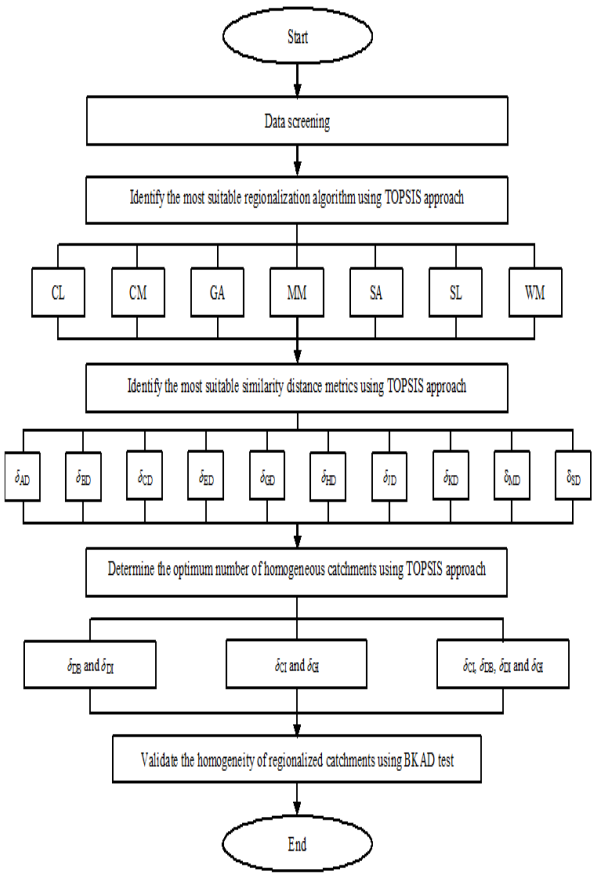

Fig. 2: Overview of the Proposed Algorithm.

3.1. Data screening

The missing daily rainfall records are imputed by applying the superior imputation algorithm of missing rainfall records which is proposed in Saeed et al. [36]. Let the imputed historical daily rainfall records for K rainfall catchments are aggregated into J months as $\hat{\mathrm{X}}=\left[\hat{x}_{jk}\right]_{JK}$; $i,(j)=1, 2, \ldots, J, (K)$. The unitization function, which is based on the coefficient of variations (as shown in Table 1), is applied to diminish the variability among the rainfall catchments:

$$\mathrm{X}_j=\frac{\hat{\mathrm{X}}_j-\min_{\hat{\mathrm{X}}_j\in\hat{\mathrm{X}}}\left\{\hat{\mathrm{X}}_j\right\}}{\max_{\hat{\mathrm{X}}_j\in\hat{\mathrm{X}}}\left\{\hat{\mathrm{X}}_j\right\}-\min_{\hat{\mathrm{X}}_j\in\hat{\mathrm{X}}}\left\{\hat{\mathrm{X}}_j\right\}} \tag{1}$$

Hence, a new transformed data, $\mathrm{X}=\left[x_{jk}\right]_{JK}$; $x_{jk}\in[0, 1]$ is resulted.

The accuracy of extreme estimates in the analysis of stationary regional frequency is liable on the inherent assumptions of no serial correlation over time and spatial independence in the rainfall catchments time series. It should be highlighted that this study is independent of these assumptions, and focuses on the non-stationary regional frequency analysis. The Mann-Kendall trend [28] test carried out on the monthly historical rainfall records showed that the rainfalls are restricted from serial correlation over time.

3.2. Agglomerative hierarchical clustering algorithms

The agglomerative hierarchical clustering (AHC) algorithm based on several mechanisms such as the complete linkage (CL), centroid (CM), group average (GA), median (MM), simple average (SA), single linkage (SL) and Ward's minimum variance (WM) is an unsupervised learning approach which is applied to identify natural homogeneous rainfall catchments. In principle, the AHC algorithm is performed with $K-1$ successive fusions, by agglomerating the closest (or farthest) pair of the rainfall catchments based on the predetermined distance metric, until K rainfall catchments is agglomerated as a single cluster (or dendrogram).

Let $\delta\left(\mathrm{X}_{p_0}, \mathrm{X}_{q_0}\right)$ represents the smallest predetermined similarity distance for a single cluster, X_p, comprising the agglomerated pair of X_{p_0} and X_{q_0} clusters. A new dendrogrammatic distance, $\delta^*\left(\mathrm{X}_p, \mathrm{X}_q\right)$, between a new single cluster, X_p, and the remaining non-agglomerated clusters, X_q, is resulted by updating the general dendrogrammatic distance function [24]:

$$\delta^*\left(\mathrm{X}_p,\mathrm{X}_q\right)=\beta_1\delta\left(\mathrm{X}_{p_0},\mathrm{X}_p\right)+\beta_2\delta\left(\mathrm{X}_{q_0},\mathrm{X}_q\right)+\beta_3\delta\left(\mathrm{X}_{p_0},\mathrm{X}_{q_0}\right)+\beta_4\left|\delta\left(\mathrm{X}_{p_0},\mathrm{X}_p\right)-\delta\left(\mathrm{X}_{q_0},\mathrm{X}_q\right)\right| \tag{2}$$

The coefficients of the distance function are β_1, β_2, β_3 and β_4, and the related formulas for the coefficients based on seven mechanisms considered in this study are summarized in Table 2. The number of rainfall catchments in cluster X_{p_0}, X_{q_0} and X_q are denoted by n_{p_0}, n_{q_0} and n_q, respectively.

Table 2: Coefficients of Distance Function for Seven Mechanisms of AHC Algorithms

Mch	Coefficients			
	β_1	β_2	β_3	β_4
CL	0.5	0.5	0	0.5
CM	$\frac{n_{p_0}}{n_{p_0}+n_{q_0}}$	$\frac{n_{p_0}}{n_{p_0}+n_{q_0}}$	$-\frac{n_{p_0}n_{q_0}}{\left(n_{p_0}+n_{q_0}\right)^2}$	0
GA	$\frac{n_{p_0}}{n_{p_0}+n_{q_0}}$	$\frac{n_{q_0}}{n_{p_0}+n_{q_0}}$	0	0
MM	0.5	0.5	-0.25	0
SA	0.5	0.5	0	0
SL	0.5	0.5	0	-0.5
WM	$\frac{n_{p_0}+n_q}{n_{p_0}+n_{q_0}+n_q}$	$\frac{n_{q_0}+n_q}{n_{p_0}+n_{q_0}+n_q}$	$-\frac{n_q}{n_{p_0}+n_{q_0}+n_q}$	0

*Note: Mch: Mechanisms

3.3. Similarity distance metrics

The AHC algorithm is frequently associated with a similarity distance (SD) metric. The Euclidean distance is a well-known SD metric, and is often applied in the AHC algorithm in previous hydrological studies. A series of comparative studies concerning the effectiveness of the SD metrics in the AHC algorithm were also carried out in other multidisciplinary areas [21], [33], [38], [41].

This study further investigates the efficiency of several SD metrics in the AHC algorithm. Ten measures of SD metrics without upper bound are considered in this study, namely the Altgower (δ_{AD}), Binomial (δ_{BD}), Canberra (δ_{CD}), Euclidean (δ_{ED}), Gower (δ_{GD}), Horn (δ_{HD}), Jaccard (δ_{JD}), Kulczynski (δ_{KD}), Manhattan (δ_{MD}) and Sørensen (δ_{SD}), which can be formulated as the followings:

$$\delta_{AD}(X_p, X_q) = \frac{1}{J - n_{\{(X_p)_j + (X_q)_j\}=0}} \sum_{j=1}^{J} \left\{ \left| (X_p)_j - (X_q)_j \right| \right\} \quad (3)$$

$$\delta_{BD}(X_p, X_q) = \sum_{j=1}^{J} \left\{ \frac{\log\left\{ \frac{\left\{2(X_p)_j\right\}^{\{(X_p)_j\}} \left\{2(X_q)_j\right\}^{\{(X_q)_j\}}}{\left\{(X_p)_j + (X_q)_j\right\}^{\{(X_p)_j + (X_q)_j\}}} \right\}}{\left\{(X_p)_j + (X_q)_j\right\}} \right\} \quad (4)$$

$$\delta_{CD}(X_p, X_q) = \frac{1}{n_{\{(X_p)_j, (X_q)_j\}>0}} \sum_{j=1}^{J} \left\{ \frac{\left| (X_p)_j - (X_q)_j \right|}{(X_p)_j + (X_q)_j} \right\} \quad (5)$$

$$\delta_{ED}(X_p, X_q) = \sqrt{\sum_{j=1}^{J} \left\{ (X_p)_j - (X_q)_j \right\}^2} \quad (6)$$

$$\delta_{GD}(X_p, X_q) = \frac{1}{J} \sum_{j=1}^{J} \left\{ \left| (X_p)_j - (X_q)_j \right| \right\} \quad (7)$$

$$\delta_{HD}(X_p, X_q) = 1 - \left\{ \frac{2\sum_{j=1}^{J}\left\{(X_p)_j\right\}\sum_{j=1}^{J}\left\{(X_q)_j\right\}\sum_{j=1}^{J}\left\{(X_p)_j (X_q)_j\right\}}{\left\{\sum_{j=1}^{J}\left\{(X_p)_j\right\}^2\left\{\sum_{j=1}^{J}\left\{(X_q)_j\right\}\right\}^2 + \left\{\sum_{j=1}^{J}\left\{(X_p)_j\right\}\right\}^2 \sum_{j=1}^{J}\left\{(X_q)_j\right\}^2\right\}} \right\} \quad (8)$$

$$\delta_{JD}(X_p, X_q) = \frac{2\sum_{j=1}^{J}\left|(X_p)_j - (X_q)_j\right|}{\sum_{j=1}^{J}\left\{(X_p)_j - (X_q)_j\right\} + \sum_{j=1}^{J}\left|(X_p)_j - (X_q)_j\right|} \quad (9)$$

$$\delta_{KD}(X_p, X_q) = 1 - \left\{ \frac{\sum_{j=1}^{J} \min_j \left\{ \left\langle (X_p)_j, (X_q)_j \right\rangle \right\}}{2\sum_{j=1}^{J}\left\{(X_p)_j\right\}} + \frac{\sum_{j=1}^{J} \min_j \left\{ \left\langle (X_p)_j, (X_q)_j \right\rangle \right\}}{2\sum_{j=1}^{J}\left\{(X_q)_j\right\}} \right\} \quad (10)$$

$$\delta_{MD}(X_p, X_q) = \sum_{j=1}^{J} \left\{ \left| (X_p)_j - (X_q)_j \right| \right\} \quad (11)$$

$$\delta_{SD}(X_p, X_q) = \frac{\sum_{j=1}^{J} \left\{ \left| (X_p)_j - (X_q)_j \right| \right\}}{\sum_{j=1}^{J} \left\{ (X_p)_j - (X_q)_j \right\}} \quad (12)$$

where $n_{\{(X_p)_j, (X_q)_j\}>0}$ is the number of non-zero monthly historical rainfall amounts in clusters X_p and $X_q; X_p, X_q \in X_k$, and $n_{\{(X_p)_j + (X_q)_j\}=0}$ is the number of pairs of zero monthly rainfall amounts in clusters X_p and X_q.

Let $\delta(X_s, X_t)$ represents the SD metric between two clusters, X_s and $X_t; X_s, X_t \in X_k$. The performance of the SD metrics in the AHC algorithms is evaluated using the cophenetic correlation coefficient (R_C), which is formulated as below:

$$R_C = \frac{\sum_{s<t} \left\{ \left\{ \delta(X_s, X_t) - \bar{\delta}(X_s, X_t) \right\} \times \left\{ \delta^*(X_s, X_t) - \bar{\delta}^*(X_s, X_t) \right\} \right\}}{\sqrt{\left\{ \sum_{s<t} \left\{ \delta(X_s, X_t) - \bar{\delta}(X_s, X_t) \right\}^2 \times \sum_{s<t} \left\{ \delta^*(X_s, X_t) - \bar{\delta}^*(X_s, X_t) \right\}^2 \right\}}} \quad (13)$$

The results of Kruskal-Wallis H test on the computed R_C indicate that there are significant differences among the ten similarity measures considered. Therefore, the R_C, which is based on similarity measures, are used as attributes for the TOPSIS algorithm to identify the most suitable AHC algorithm.

3.4. Multi-criteria decision making algorithm

In previous studies, the AHC algorithms are often determined based on prior knowledge [15], [17], [32]. In other studies, Pansera et al. [35] and Saraçli et al. [38] proposed the identification of the most appropriate AHC algorithm based on the R_C. Previous studies [2], [35] suggested that the identification of the optimum number of homogeneous catchments is based on several internal clustering validation indices. Therefore, the TOPSIS algorithm, which is a multi-criteria decision making algorithm originated from Hwang and Yoon [19], is suggested in this study as an alternative algorithm to identify the most suitable AHC algorithm and to determine the optimum number of homogenous rainfall catchments.

The TOPSIS algorithm is an effective probabilistic analytical model, which is extensively researched in the literatures of multi-disciplinary areas [12], [34,] [37], [40]. Let $T = \left[t_{pq}\right]_{PQ}$; $p,(q) = 1, 2, \cdots, P,(Q)$ represents the normalized matrix of R_C with P attributes and Q alternatives. The main objective of the normalized matrix is to ensure the comparability across attributes.

Based on the principle of TOPSIS algorithm, the designated best alternative shows the highest value of relative closeness (C), and is given as:

$$C = \max_q \left\{ \frac{\sqrt{\sum_{p=1}^{P}\left(\theta_{pq} - \theta_q^-\right)^2}}{\sqrt{\sum_{p=1}^{P}\left(\theta_{pq} - \theta_q^+\right)^2} + \sqrt{\sum_{p=1}^{P}\left(\theta_{pq} - \theta_q^-\right)^2}} \right\} \quad (14)$$

where $\theta_q^+ = \max_q\left\{\theta_{pq}\right\}$, $\theta_q^- = \min_q\left\{\theta_{pq}\right\}$, $0 \le C \le 1$, and $\theta_{pq} = w_q t_{pq}$ is the weighted normalized observations with weight function:

$$w_q = \frac{\sum_{p=1}^{P}\left(t_{pq} - \bar{t}_q\right)}{\sum_{q=1}^{Q}\left(\sum_{p=1}^{P}\left(t_{pq} - \bar{t}_q\right)\right)} \quad (15)$$

on condition that $\sum_{q=1}^{Q} w_q = 1$.

3.5. Homogeneity validation indices

In principle, clustering validation indices can be categorized into external and internal criteria, and are used to evaluate the goodness of the identified natural homogenous cluster [29]. Since external clustering validation indices requires prior information regarding the data, several hydrological studies [2], [22], [35] used internal validation indices in identifying the optimum number of homogeneous catchments. However, past studies have shown that the determination of the optimum number of homogeneous catchments based on selected validation indices are considerably subjective, due to inappropriate combinations of validation indices

that may affect the final results [8], [13]. This uncertainty, however, may be overcome by using several well-known validation indices, such as the C-index (δ_{CI}), Davies-Bouldin index (δ_{DB}), Dunn index (δ_{DI}), and Gamma index (δ_{GI}), which are suggested in this study as attributes for the TOPSIS algorithm.

Suppose $\mathrm{X}=\left[x_{jk}\right]_{JK}$ are partitioned into M disjoint clusters of homogeneous catchments, $\mathrm{X}=\left[\mathrm{C}_i\right]_{JK_i}; i=1, 2, \ldots, M$ with centroid $\bar{\mathrm{C}}_i=\sum_{x_{jk_i}\in \mathrm{C}_i}\frac{x_{jk_i}}{JK_i}$. Therefore, the aforementioned validation indices can be expressed as:

$$\delta_{\text{CI}}=\frac{\left\{\sum_{\mathrm{C}_i\in\mathrm{X}}\sum_{x_{1jk},x_{2jk}\in\mathrm{C}_i}\left\{\delta_{\text{ED}}\left(x_{1jk},x_{2jk}\right)\right\}-\sum_{x_{\varphi_1 jk},x_{\varphi_2 jk}\in\mathrm{X}}\left\{\min_{\varphi_1\neq\varphi_2}\left\{\delta\left(x_{\varphi_1 jk},x_{\varphi_2 jk}\right)\right\}\right\}\right\}}{\left\{\sum_{x_{\varphi_1 jk},x_{\varphi_2 jk}\in\mathrm{X}}\left\{\max_{\varphi_1\neq\varphi_2}\left\{\delta\left(x_{\varphi_1 jk},x_{\varphi_2 jk}\right)\right\}\right\}-\sum_{x_{\varphi_1 jk},x_{\varphi_2 jk}\in\mathrm{X}}\left\{\min_{\varphi_1\neq\varphi_2}\left\{\delta\left(x_{\varphi_1 jk},x_{\varphi_2 jk}\right)\right\}\right\}\right\}} \tag{16}$$

$$\delta_{\text{DB}}=\frac{\sum_{\mathrm{C}_{i_a},\mathrm{C}_{i_b}\in\mathrm{X}}\left\{\max_{i_a\neq i_b}\left\{\frac{\left\{\sum_{x_{jk}\in\mathrm{C}_{i_a}}\left\{\frac{\delta_{\text{ED}}\left(x_{jk},\bar{\mathrm{C}}_{i_a}\right)}{JK_{i_a}}\right\}+\sum_{x_{jk}\in\mathrm{C}_{i_b}}\left\{\frac{\delta_{\text{ED}}\left(x_{jk},\bar{\mathrm{C}}_{i_b}\right)}{JK_{i_b}}\right\}\right\}}{\delta_{\text{ED}}\left(\bar{\mathrm{C}}_{i_a},\bar{\mathrm{C}}_{i_b}\right)}\right\}\right\}}{M} \tag{17}$$

$$\delta_{\text{DI}}=\frac{\min_{\mathrm{C}_{i_A},\mathrm{C}_{i_B}\in\mathrm{X}}\left\{\min_{i_A\neq i_B}\left\{\delta_{\text{ED}}\left(x_{ij}\in\mathrm{C}_{i_A},x_{ij}\in\mathrm{C}_{i_B}\right)\right\}\right\}}{\max_{\mathrm{C}_i\in\mathrm{X}}\left\{\max_{x_{1ij},x_{2ij}\in\mathrm{C}_i}\left\{\delta_{\text{ED}}\left(x_{1ij},x_{2ij}\right)\right\}\right\}} \tag{18}$$

$$\delta_{\text{GI}}=\frac{\sum_{x_{\varphi_1 jk},x_{\varphi_2 jk}\in\mathrm{X}}\sum_{x_{1jk},x_{2jk}\in\mathrm{C}_i}I\left\{\delta_{\text{ED}}\left(x_{\varphi_1 jk},x_{\varphi_2 jk}\right)<\delta_{\text{ED}}\left(x_{1jk},x_{2jk}\right)\right\}}{\left(\sum_{\mathrm{C}_i\in\mathrm{X}}\left\{\binom{JK_i}{2}\right\}\right)\left(\binom{JK}{2}-\sum_{\mathrm{C}_i\in\mathrm{X}}\left\{\binom{JK_i}{2}\right\}\right)} \tag{19}$$

3.6. Bootstrap K-sample Anderson darling test

In the past decades, several homogeneity tests [9], [27], [46], [48] were introduced in hydrological literatures. These homogeneity tests were routinely applied to validate regionalized homogeneous catchments, especially after the introduction of the L-moment heterogeneity measures by Hosking and Wallis [18]. Even though the efficiency of the L-moment heterogeneity measures is restricted to highly skewed hydrological data [45], this restriction can be competently overcome by using the Bootstrap K-sample Anderson Darling (BKAD) test.

The BKAD test is a generalized classical Anderson Darling goodness-of-fit test which is free from any statistical assumption [39],]45]. Let $x_{(1)}<x_{(2)}<\cdots<x_{(JK)}$ be the pooled ordered sample of X. The test statistic of BKAD, ϕ^2_{BKAD} and its variance, $\mathrm{V}\left(\phi^2_{KAD}\right)$, can be formulated as:

$$\phi^2_{KAD}=\sum_{j=1}^{JK-1}\frac{\left(n_{jk}K-j\right)^2}{j\left(JK-j\right)} \tag{20}$$

$$\mathrm{V}\left(\phi^2_{BKAD}\right)=\frac{\Gamma\left(JK-3\right)}{\Gamma\left(JK\right)}\left(\gamma_1+\gamma_2 JK+\gamma_3\left(JK\right)^2+\gamma_4\left(JK\right)^3\right) \tag{21}$$

where n_{jk} represents the number of observations in the kth sample that are no more than the jth smallest observation in the pooled samples, and $\gamma_1, \gamma_2, \gamma_3$ and γ_4 are defined as:

$$\begin{aligned}
\gamma_1&=2K\left(3K+\sum_{j=1}^{JK-1}\frac{\left(K-2\right)}{j}\right)\\
\gamma_2&=-\frac{2}{J}\left(JK^2-3JK+3K-\sum_{j=1}^{JK-1}\frac{\left(3JK^2+2JK+2J+K\right)}{j}-\sum_{j=1}^{JK-2}\frac{JK}{\left(JK-j\right)\left(j+1\right)}\right)\\
\gamma_3&=-\frac{2}{J}\left(2JK^2+3J+3K-\sum_{j=1}^{JK-1}\frac{\left(4JK-4J-7K\right)}{j}-\sum_{j=1}^{JK-2}\frac{JK^2+K+2J}{\left(JK-j\right)\left(j+1\right)}\right)\\
\gamma_4&=-\frac{2}{J}\left(3JK+3J+5K-\sum_{j=1}^{JK-2}\frac{2JK-2J-3K}{\left(JK-j\right)\left(j+1\right)}\right)
\end{aligned} \tag{22}$$

Based on this rank test, the regionalized homogeneous catchments have considerable heterogeneity if and only if

$$T_{BKAD}=\frac{\theta^2_{BKAD}-\left(K-1\right)}{\mathrm{V}\left(\lambda^2_{BKAD}\right)}\geq T_{K-1,\,\alpha} \tag{23}$$

Since the BKAD test is performed solely based on the ranks of sample observations, the stability property of this rank test is questionable. As an alternative, the bootstrap resampling approach is applied to determine the acceptable limits of the BKAD test.

4. Results and discussion

4.1. Identification of homogeneous rainfall catchments

Fig. 3 illustrates the descriptive statistics of the monthly rainfall historical records for 20 selected rainfall catchments in Kuantan River Basin after applying the unitization function. Based on the measure of central tendency, the average monthly rainfalls of Station 18 is significantly higher, while the average monthly rainfalls of Station 01 is significantly lower compared to other rainfall catchments. The non-parametric multiple comparison test show that there are significant differences between Station 05 with Stations 02 and 08. The statistical evidences indicate that the average monthly rainfalls of Station 14 is higher than Stations 02, 03, 06, 08 and 18 at $\alpha=0.05$.

Table 3 and Table 4 respectively illustrate the performances of seven AHC algorithms and ten selected SD metrics based on the TOPSIS algorithm. The GA algorithm (shown in Table 3) is more superior than the other six algorithms, as displayed by the highest relative closeness. In terms of attributes, the δ_{AD} (shown in Table 4) is more superior than the other nine SD metrics. Thus, the GA algorithm with δ_{AD} attribute is the most suitable algorithm among the seventy AHC algorithms tested in this study. It should be noted that seven AHC algorithms with ten SD metrics attributes considered in this study resulted into a total of seventy AHC algorithms.

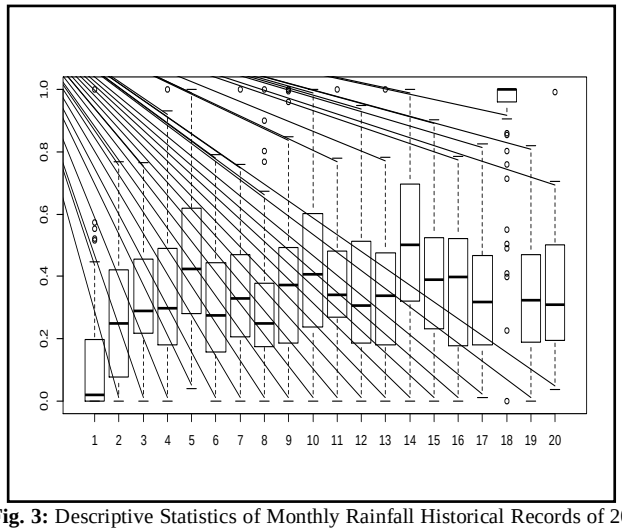

Fig. 3: Descriptive Statistics of Monthly Rainfall Historical Records of 20 Selected Rainfall Catchments in Kuantan River Basin.

Table 3: Ranking and Performance for AHC Algorithms

SD	AHC						
	CL	CM	GA	MM	SA	SL	WM
δ_{AD}	0.9621	0.9646	0.972	0.9574	0.9649	0.9690	0.9298
δ_{BD}	0.9099	0.9576	0.9669	0.9550	0.9578	0.9583	0.9110
δ_{CD}	0.8984	0.9407	0.9566	0.9313	0.9525	0.9426	0.8851
δ_{ED}	0.9501	0.9554	0.9672	0.9254	0.9590	0.9630	0.9194
δ_{GD}	0.9625	0.9636	0.9712	0.9536	0.9632	0.9668	0.9032
δ_{HD}	0.8542	0.9352	0.9383	0.8917	0.9204	0.9336	0.8121
δ_{JD}	0.9089	0.9161	0.9438	0.9056	0.9351	0.9268	0.8379
δ_{KD}	0.8643	0.8853	0.9117	0.8305	0.9005	0.8967	0.7516
δ_{SD}	0.9265	0.9402	0.9499	0.9280	0.9422	0.9359	0.8830
δ_{MD}	0.9625	0.9636	0.9712	0.9536	0.9632	0.9668	0.9032
Ranked	5	4	1	6	3	2	7

Table 4: Ranking and Performance for SD Metrics

SD	AHC							Ranked
	CL	CM	GA	MM	SA	SL	WM	
δ_{AD}	0.9621	0.9646	0.9720	0.9574	0.9649	0.9690	0.9298	1
δ_{BD}	0.9099	0.9576	0.9669	0.9550	0.9578	0.9583	0.9110	5
δ_{CD}	0.8984	0.9407	0.9566	0.9313	0.9525	0.9426	0.8851	7
δ_{ED}	0.9501	0.9554	0.9672	0.9254	0.9590	0.9630	0.9194	2
δ_{GD}	0.9625	0.9636	0.9712	0.9536	0.9632	0.9668	0.9032	3
δ_{HD}	0.8542	0.9352	0.9383	0.8917	0.9204	0.9336	0.8121	9
δ_{JD}	0.9089	0.9161	0.9438	0.9056	0.9351	0.9268	0.8379	8
δ_{KD}	0.8643	0.8853	0.9117	0.8305	0.9005	0.8967	0.7516	10
δ_{SD}	0.9265	0.9402	0.9499	0.9280	0.9422	0.9359	0.8830	6
δ_{MD}	0.9625	0.9636	0.9712	0.9536	0.9632	0.9668	0.9032	3

4.2. Identification of optimum number of homogeneous rainfall catchments

An approach to determine the optimum number of homogeneous catchments using the best algorithm, which is the group average (GA) algorithm with the Altgower (δ_{AD}) attribute, is presented in this section.

The Davies-Bouldin index (δ_{DB}) and the Dunn index (δ_{DI}) are the two well-known internal clustering validation indices applied in multidisciplinary studies [2], [4], [23], [30], [35]. Since the value of δ_{DB} should be as low as possible, and the value of δ_{DI} should be as high as possible, the absolute difference between δ_{DB} and δ_{DI} are used as a comparative baseline in this study to determine the optimum number of homogeneous rainfall. This comparative baseline is shown in Table 5.

The performance of several combinations of internal clustering validation indices, which are used as attributes for the TOPSIS algorithm, is illustrated in the Table 6. Four indices, namely the C-index (δ_{CI}), Davies-Bouldin index (δ_{DB}), Dunn index (δ_{DI}), and Gamma index (δ_{GI}), and four differences approaches, namely $|\delta_{DB}-\delta_{DI}|^1$, $\delta_{DB},\delta_{DI}{}^2$, $\delta_{CI},\delta_{GI}{}^3$, and $\delta_{DB},\delta_{DI}\delta_{CI},\delta_{GI}{}^4$, are considered in this study. The results of the Mann-Whitney U-test on the attributes of approach[2] and approach[3] indicate that there are significant

differences among the internal clustering validation indices, respectively. Therefore, the combination of these all four internal clustering validation indices is also considered.

Table 5: Performance for Homogeneity Validation Indices

Number of clusters	Homogeneity validation indices			
	δ_{DB}	δ_{DI}	δ_{CI}	δ_{GI}
2	0.7004	0.4435	0.3281	0.6639
3	1.7628	0.3699	0.2617	0.6959
4	1.7690	0.2896	0.2682	0.5016
5	1.4409	0.3187	0.2697	0.5890
6	1.4875	0.3105	0.2548	0.6348
7	1.2442	0.5863	0.5153	0.7374
8	1.2026	0.5863	0.5092	0.6912
9	1.1021	0.6232	0.5280	0.7448
10	0.9941	0.6232	0.5282	0.7314
11	0.9289	0.6232	0.4938	0.7581
12	0.8452	0.6232	0.4733	0.7862
13	0.7791	0.6232	0.4810	0.7650
14	0.7836	0.6232	0.4797	0.6827
15	0.6455	0.8466	0.5806	0.9583
16	0.5428	0.8466	0.5226	0.9632
17	0.4494	0.9902	0.6163	0.9973
18	0.4350	0.7969	0.6332	0.9840
19	0.2907	1.0423	1.0000	1.0000

Table 6: Ranking for Homogeneity Validation Indices

Number of clusters	Ranked			
	$\lvert\delta_{DB}-\delta_{DI}\rvert^{1}$	$\delta_{DB},\delta_{DI}{}^{2}$	$\delta_{CI},\delta_{GI}{}^{3}$	$\delta_{DB},\delta_{DI},\delta_{CI},\delta_{GI}{}^{4}$
2	18	17	18	18
3	10	14	17	14
4	12	12	10	12
5	7	2	8	1
6	6	1	13	2
7	1	7	14	7
8	2	8	7	8
9	3	9	3	9
10	9	3	11	3
11	8	4	15	4
12	5	5	16	6
13	4	6	6	5
14	11	10	5	10
15	13	11	2	11
16	14	13	4	13
17	15	15	9	16
18	16	16	12	17
19	17	18	1	15

Fig. 4 presents the results of regionalized homogeneous catchments based on the most appropriate AHC algorithm (GA algorithm with δ_{AD} attribute) and four differences approaches (approach[1], approach[2], approach[3], and approach[4]). The optimum number of homogeneous catchments, which is provided by approach[1], is seven clusters. It can also be seen from the figures that the optimum number of homogeneous catchments reduce to six and five clusters when δ_{DB}, δ_{DI} (from approach[2]) and δ_{DB}, δ_{DI}, δ_{CI}, δ_{GI} (from approach[4]) are used as attributes, but increase to nineteen clusters when δ_{CI} and δ_{GI} (in approach[3]) are used as attributes. The optimum number of homogeneous catchments suggested in this study (seven clusters) is reasonable as it is consistent with the results of the non-parametric multiple comparison tests.

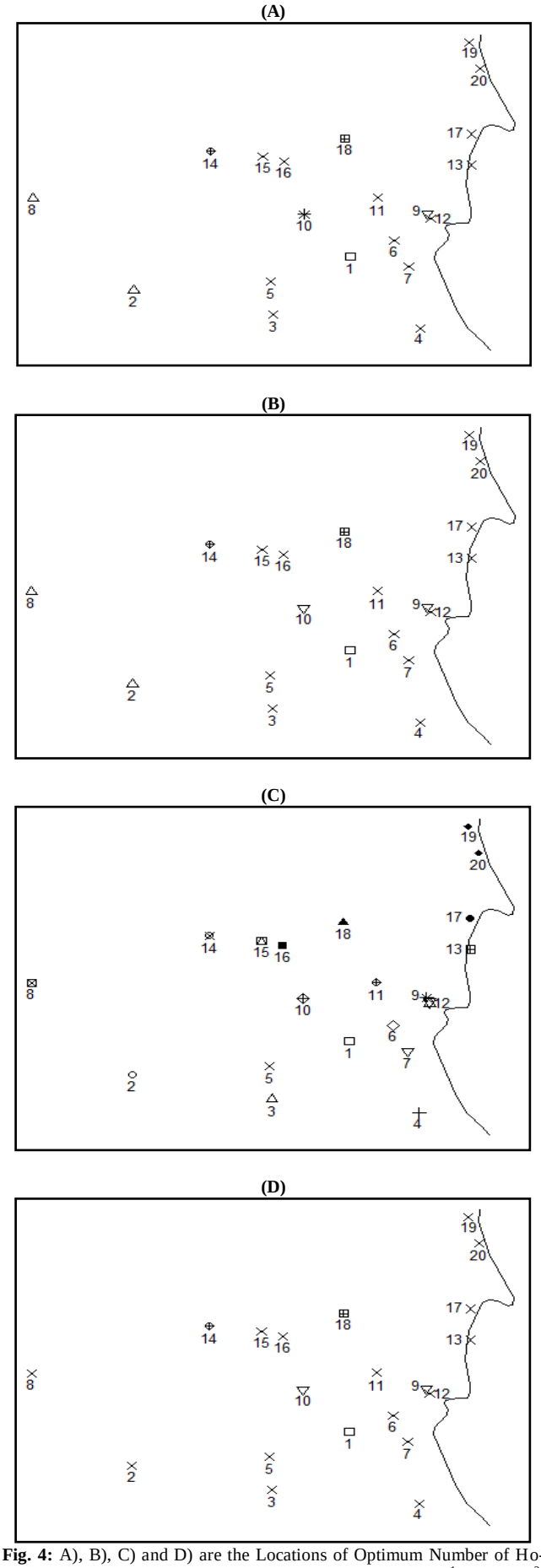

Fig. 4: A), B), C) and D) are the Locations of Optimum Number of Homogeneous Rainfall Catchments Determined from approach[1], approach[2], approach[3] and approach[4], Respectively.

4.3. Validation of homogeneous regionalized rainfall catchments

The homogeneity of regionalized rainfall catchments in this study is suggested to be validated by the BKAD test. The results of the BKAD test with sample size 50, 100, 500, 1000, 5000 and 10000 are presented in Table 7. Several algorithms from previous studies are also compared, namely Modified Ahmad et al. [2][1], Modified Hamdan et al. [17][1], Modified Ahmad et al. [2][2,4] and Modified Hamdan et al. [17][2,4]. The results in Table 7 show that all regionalized rainfall catchments are significantly homogeneous at $\alpha = 0.05$, and are invariant to the number of Bootstrap samples. However, the statistical significance do not imply that all of the algorithms are efficient and applicable in this study. Several algorithms, however, have resulted in a large number of clusters such as Modified Ahmad et al. [2][1] and Hamdan et al. [17][1]. Even though the results from Modified Ahmad et al. [2][2] and Hamdan et al. [17][2] algorithms are not shown here, these algorithms have resulted in nineteen homogeneous clusters.

The results of non-parametric multiple comparison tests indicate that Stations 01, 14 and 18 are outliers as the three rainfall catchments cannot be merged with other rainfall catchments. The non-parametric tests also show that Stations 02 and 08 should be not located in the same cluster with Station 05. Therefore, the results from the approach[4] algorithm are considered as inappropriate since they are inconsistent with the results of the non-parametric multiple comparison tests.

Table 7: Validation of Homogeneous Regionalized Rainfall Catchments Using BKAD Test with Various Sample Sizes

Algorithm	Cluster	Homogeneous rainfall catchments	P-values					
			50	100	500	1000	5000	10000
Modified Ahmad et al. [2][1]	1	01	-	-	-	-	-	-
	2	02, 08	0.6200	0.6100	0.5320	0.5220	0.5008	0.4946
	3	03, 06, 12	0.6000	0.5500	0.5800	0.5800	0.5870	0.5880
	4	04, 07, 15, 16	0.7600	0.7200	0.6680	0.6690	0.6844	0.6898
	5	05, 11	0.4400	0.3700	0.3980	0.3880	0.3950	0.3997
Modified Hamdan et al. [17][1]	6	09	-	-	-	-	-	-
	7	10	-	-	-	-	-	-
	8	13	-	-	-	-	-	-
	9	14	-	-	-	-	-	-
	10	17, 19, 20	0.1200	0.0700	0.0600	0.0620	0.0792	0.0796
	11	18	-	-	-	-	-	-
Modified Ahmad et al. [2][2,4]	1	01	-	-	-	-	-	-
	2	02, 08, 09, 10	0.4200	0.4000	0.3480	0.3470	0.3336	0.3321
Modified Hamdan et al. [17][2,4]	3	03, 06, 12, 13, 17, 19, 20	0.8400	0.8400	0.8200	0.8260	0.8228	0.8279
	4	04, 05, 07, 11, 15, 16	0.6600	0.7400	0.5940	0.6100	0.6342	0.6210
	5	14	-	-	-	-	-	-
	6	18	-	-	-	-	-	-
approach[1]	1	01	-	-	-	-	-	-
	2	02, 08	0.6200	0.6100	0.5320	0.5220	0.5008	0.4946
	3	03, 04, 05, 06, 07, 11, 12, 13, 15, 16, 17, 19, 20	0.9600	0.9300	0.9280	0.9240	0.9252	0.9306
	4	09	-	-	-	-	-	-
	5	10	-	-	-	-	-	-
	6	14	-	-	-	-	-	-
	7	18	-	-	-	-	-	-
approach[2]	1	01	-	-	-	-	-	-
	2	02, 08	0.6200	0.6100	0.5320	0.5220	0.5008	0.4946
	3	03, 04, 05, 06. 07, 11, 12, 13, 15, 16, 17, 19, 20	0.6000	0.4900	0.5140	0.5090	0.5094	0.5075
	4	09, 10	0.9200	0.9900	0.9260	0.9190	0.9292	0.9297
	5	14	-	-	-	-	-	-
	6	18	-	-	-	-	-	-
approach[4]	1	01	-	-	-	-	-	-
	2	02, 03, 04, 05, 06, 07, 08, 11, 12, 13, 15, 16, 17, 19, 20	0.9600	0.9700	0.9260	0.9190	0.9190	0.9184
	3	09, 10	0.6200	0.5700	0.4780	0.4950	0.5098	0.5045
	4	14	-	-	-	-	-	-
	5	18	-	-	-	-	-	-

The algorithms that produce smaller and suitable number of homogeneous clusters are Modified Ahmad et al. [2][2,4], Hamdan et al. [17][2,4], approach[1] and approach[2]. By comparing these four algorithms, the approach[1] algorithm is the most appropriate algorithm to regionalize the twenty rainfall catchments considered in this study. The optimum number of homogenous catchments suggested by this algorithm is consistent with the results of the non-parametric multiple comparison tests. However, under this algorithm, Stations 09 and 10 are misplaced because their average monthly rainfall amounts are not significantly different compared to all rainfall catchments in Cluster 3.

After merging Clusters 3 and 4 together, the BKAD tests show very strong statistical evidences of homogenous rainfall catchments, where the p-values for sample size 50, 100, 500, 1000, 5000 and 10000 are 0.9200, 0.9300, 0.8920, 0.8850, 0.8932 and 0.8949, respectively. After excluding the outliers (Stations 01, 14, and 18), the twenty rainfall catchments can be finally regionalized into two different homogeneous rainfall catchments, which are Cluster 1 (Stations 02, 03, 04, 05, 06, 07, 08, and 11) and Cluster 2 (Stations 09, 10, 12, 13, 15, 16, 17, 19, and 20).

5. Conclusion

This study has proposed a new regionalization algorithm, which does not require prior knowledge and capabilities, to determine the most suitable agglomerative hierarchical clustering (AHC) algorithm and the optimum number of homogeneous rainfall catchments for non-stationary rainfall time series. The new algorithm is based on the TOPSIS algorithm, and is used for regionalizing homogeneous rainfall catchments from twenty selected monthly rainfall time series of monitoring stations in Kuantan River Basin, Malaysia. The regionalized homogeneous catchments resulted from the proposed algorithms are suggested to be validated using the Bootstrap K-sample Anderson-Darling (BKAD) test with various sample sizes. The results show that the group average (GA) agglomerative hierarchical clustering (AHC) algorithm with the Altgower (δ_{AD}) similarity metrics is the best algorithm out of a total of seventy AHC algorithms considered. The results also indicate that the BKAD test is invariant towards the number of Boot-

strap samples in the validation of homogeneous rainfall catchments. For future works, the proposed regionalized algorithms are suggested to be implemented to the non-stationary rainfall time-series from East-Coast regions, Malaysia.

Acknowledgement

The authors would like to thank the Department of Irrigation and Drainage Malaysia for providing the data of this research work. The authors also would like to acknowledge the Universiti Malaysia Pahang (UMP) for providing the flagship research grant RDU150393 and the internal research grant RDU1703184.

References

[1] V. Agilan, N.V. Umamahesh, Is the covariate based non-stationary rainfall IDF curve capable of encompassing future rainfall changes, Journal of Hydrology 541(B) (2016) 1441-1455.

[2] N.H. Ahmad, I.R. Othman, S.M. Deni, Hierarchical cluster approach for regionalization of Peninsular Malaysia based on the precipitation amount, Proceedings of the International Conference on Science & Engineering in Mathematics, Chemistry and Physics (2013), https://doi.org/10.1088/1742-6596/423/1/012018.

[3] M.M. Alam, G. Morshed, C. Siwar, M.W. Murad, Initiatives and challenges of agricultural crop sector in East Coast Economic Region (ECER) development projects in Malaysia, American-Eurasian Journal Agriculture & Environmental Sciences 12(7) (2012) 922-931.

[4] N. Anuar, Z. Zakaria, Electricity load profile determination by using fuzzy C-Means and probability neural network, Energy Procedia 14 (2012) 1861-1869. https://doi.org/10.1016/j.egypro.2011.12.1180.

[5] P.A. Baeriswyl, M. Rebetez, Regionalisation of precipitation in Switzerland by means of principal component analysis, Theoretical and Applied Climatology 58(1-2) (1997) 31-41. https://doi.org/10.1007/BF00867430.

[6] D.H. Burn, Z. Zrinji, M. Kowalchuk, Regionalization of catchments for regional flood frequency analysis, Journal Hydrologic Engineering 2(2) (1997) 76-82. https://doi.org/10.1061/(ASCE)1084-0699(1997)2:2(76).

[7] G.S. Cavadias, T.B.M.J. Ouarda, B. Bobée, C. Girard, A canonical correlation approach to the determination of homogeneous regions for regional flood estimation of ungauged basins, Hydrological Sciences Journal 46(4) (2001) 499-511. https://doi.org/10.1080/02626660109492846.

[8] M. Charrad, N. Ghazzali, V. Boiteau, A. Niknafs, NbClust: an R package for determining the relevant number of clusters in a data set, Journal of Statistical Software 61(6) (2014), https://doi.org/10.18637/jss.v061.i06.

[9] J.U. Chowdhury, J.R. Stedinger, L-H. Lu, Goodness-of-fit tests for regional generalized extreme value flood distributions, Water Resources Research 27(7) (1991) 1765-1776. https://doi.org/10.1029/91WR00077.

[10] Z.L. Chuan, N. Ismail, W.L. Shinyie, T.L. Ken, S.-F. Fam, A. Senawi, W.N.S.W. Yusoff, The efficiency of average linkage hierarchical clustering algorithm associated multi-scale bootstrap resampling in identifying homogeneous precipitation catchments, IOP Conference Series: Materials Science and Engineering 342 (2018) 012070, https://doi.org/10.1088/1757-899X/342/1/012070.

[11] P.S.P. Cowpertwait, A regionalization method based on a cluster probability model, Water Resources Research 47(11) (2011) W11525, https://doi.org/10.1029/2011WR011084.

[12] H. Deng, C.H. Yeh, R.J. Willis, Inter-company comparison using modified TOPSIS with objective weights, Computers and Operations Research 27(10) (2000) 963-973. https://doi.org/10.1016/S0305-0548(99)00069-6.

[13] A. Dudek Cluster quality indexes for symbolic classification-an examination, In: Decker R, Lenz H-J (ed) Advances in Data Analysis, Springer, Heidelberg, 2007. https://doi.org/10.1007/978-3-540-70981-7_4.

[14] S-F. Fam, A.A. Jemain, W.Z.W. Zin, Spatial analysis of socioeconomic deprivation in Peninsular Malaysia, International Journal of Arts & Sciences 4(17) 241-255.

[15] N.B. Guttman, The use of L-moments in the determination of regional precipitation climates, Journal of Climate 13 (1993) 547-566. https://doi.org/10.1175/1520-0442(1993)006<2309:TUOLMI>2.0.CO;2.

[16] M.J. Hall, A.W. Minns, A.K.M. Ashrafuzzaman, The application of data mining techniques for the regionalisation of hydrological variables, Hydrology and Earth System Sciences 6(4) (2002) 685-694. https://doi.org/10.5194/hess-6-685-2002.

[17] M.F. Hamdan, J. Suhaila, A.A. Jemain, Clustering rainfall pattern in Malaysia using functional data analysis, AIP Conference Proceedings 1643(1) (2015) 349-355. https://doi.org/10.1063/1.4907466.

[18] J.R.M. Hosking, J.R. Wallis, Some statistics useful in regional frequency analysis, Water Resources Research 29(2) (1993) 271-281. https://doi.org/10.1029/92WR01980.

[19] C.L. Hwang, K. Yoon, Multiple attribute decision making methods and applications a state-art-of-the-art survey, Springer-Verlag, Heidelberg, 1981.

[20] R. Jackson, Occupy World Street: A global roadmap for radical economic and political reform, Chelsea green, Hartford, 2012.

[21] P.A. Jaskowiak, R.J. Campello, I.G. Costa, on the selection of appropriate distances for gene expression data clustering, BMC Informatics 15 (2014) https://doi.org/10.1186/1471-2105-15-S2-S2.

[22] S. Kannan, S. Ghosh, Prediction of daily rainfall state in a river basin using statistical downscaling from GCM output. Stochastic Environmental Research and Risk Assessment 25(4) (2011) 457-474. https://doi.org/10.1007/s00477-010-0415-y.

[23] J. Kianfar, P. Edara, A data mining approach to creating fundamental traffic flow diagram, Procedia Social and Behavioral Sciences 104 (2013) 430-439. https://doi.org/10.1016/j.sbspro.2013.11.136.

[24] G.N. Lance, W.T. Williams, A general theory of classificatory sorting strategies 1. Hierarchical systems, The Computer Journal 9(4) (1967) 373-380. https://doi.org/10.1093/comjnl/9.4.373.

[25] H. Li, J. Sun, H. Zhang, J. Zhang, K. Jung, J. Kim, Y, Xuan, X. Wang, F. Li, What Large Sample Size Is Sufficient for Hydrologic Frequency Analysis?—A Rational Argument for a 30-Year Hydrologic Sample Size in Water Resources Management, Water 10(4) (2018) 430, https://doi.org/10.3390/w10040430.

[26] G-F. Lin, L-H. Chen, Identification of homogeneous regions for regional frequency analysis using the self-organizing map, Journal of Hydrology 324(1-4) (2006) 1-9. https://doi.org/10.1016/j.jhydrol.2005.09.009.

[27] L-H. Lu, J.R. Stedinger, Sampling variance of normalized GEV/PWM quantile estimators and a regional homogeneity test, Journal of Hydrology 138(1-2) (1992) 223-245. https://doi.org/10.1016/0022-1694(92)90166-S.

[28] H.B. Mann, Nonparametric tests against trend, Econometrica 13(3) (1945) 245-259. https://doi.org/10.2307/1907187.

[29] U. Maulik, S. Bandyopadhyay, Performance evaluation of some clustering algorithms and validity indices, IEEE Transactions on Pattern Analysis and Machine Intelligence 24(12) (2002) 1650-1654. https://doi.org/10.1109/TPAMI.2002.1114856.

[30] S.P. Mishra, D. Mishra, S. Patnaik, An integrated robust semi-supervised framework for improving cluster reliability using ensemble method for heterogeneous datasets, Karbala International Journal of Modern Science 1(4) (2015) 200-211. https://doi.org/10.1016/j.kijoms.2015.11.004.

[31] M.F.M, Nasir, M.A. Zali, H. Juahir, H. Hussain, S.M. Zain, N. Ramli, Application of receptor models on water quality data in source apportionment in Kuantan River Basin, Iranian Journal of Environmental Health Science & Engineering 9(1) (2012). https://doi.org/10.1186/1735-2746-9-18.

[32] C.S. Ngongondo, C-Y. Xu, L.M. Tallaksen, B. Alemaw, T. Chirwa, Regional frequency analysis of rainfall extremes in Southern Malawi using the index rainfall and L-moments approaches, Stochastic Environmental Research and Risk Assessment 25(7) (2011) 939-955. https://doi.org/10.1007/s00477-011-0480-x.

[33] D.T. Nguyen, Clustering with multiviewpoint-based similarity measure, IEEE Transactions on Knowledge and Data Engineering 24(6) (2012) 988-1001. https://doi.org/10.1109/TKDE.2011.86.

[34] S. Opricovic, G-H. Tzeng, Comprise solution by MCDM methods: a comparative analysis of VIKOR and TOPSIS, European Journal of Operational Research 156(2) (2004) 445-455. https://doi.org/10.1016/S0377-2217(03)00020-1.

[35] W.A. Pansera, B.M. Gomes, M.A.V. Boas, E.Ld. Mello, Clustering rainfall stations aiming regional frequency analysis, Journal of Food, Agriculture & Environment 11(2) (2013) 877-885.

[36] G.A.A. Saeed, Z.L. Chuan, R. Zakaria, W.N.S.W. Yusoff, M.Z. Salleh, Determination of the best single imputation algorithm for missing rainfall data treatment, Journal of Quality Measurement and Analysis 12(1-2) (2016) 79-87.

[37] H. Safari, E. Khanmohammadi, A. Hafezamini, S.S. Ahangari, A new technique for multi criteria decision making based on modified

similarity method, Middle-East Journal of Scientific Research 14(5) (2013) 712-719.
[38] S. Saraçli, N. Doğan, İ. Doğan, Comparison of hierarchical cluster analysis methods by cophenetic correlation, Journal of Inequalities and Applications 2013(203) (2013).
[39] F.W. Scholz, M.A. Stephens, K-sample Anderson-Darling Tests, Journal of American Statistical Association 82(399) (1987) 918-924.
[40] H-S. Shih, H-J. Shyur, E.S. Lee, An extension of TOPSIS for group decision-making, Mathematical and Computer Modelling 45(7-8) (2007) 801-813. https://doi.org/10.1016/j.mcm.2006.03.023.
[41] A.S. Shirkhorshidi, S. Aghabozorgi, T.Y. Wah, A comparison study on similarity and dissimilarity measures in clustering continuous data, PLoS One 10(12) (2015). https://doi.org/10.1371/journal.pone.0144059.
[42] K.K. Singh, S.V. Singh, Space-time variation and regionalization of seasonal and monthly summer monsoon rainfall on sub-Himalayan region and Gangetic plains of India, Climate Research 6(3) (1996) 251-262. https://doi.org/10.3354/cr006251.
[43] F.T. Tangang, L. Juneng, E. Salimun, K.M. Sei, L.J. Le, H. Muhamad, Climate change and variability over Malaysia: Gaps in science and research information, Sains Malaysiana, 41(11) (2012) 1355-1366.
[44] B. Venkatesh, M.K. Jose, Identification of homogeneous rainfall regimes in parts of Western Ghats region of Karnataka, Journal of Earth System Science 116(4) (2007) 321-329. https://doi.org/10.1007/s12040-007-0029-z.
[45] A. Viglione, F. Laio, P. Claps, A comparison of homogeneity tests for regional frequency analysis, Water Resources Research 43 (2007) W03428, https://doi.org/10.1029/2006WR005095.
[46] S.E. Wiltshire, Regional flood frequency analysis I: Homogeneity statistics, Hydrological Sciences Journal 31(3) (1986a) 321-333. https://doi.org/10.1080/02626668609491051.
[47] S.E. Wiltshire, Regional flood frequency analysis II: Multivariate classification of drainage basins in Britain, Hydrological Sciences Journal 31(3) (1986b) 335-346. https://doi.org/10.1080/02626668609491052.
[48] S.E. Wiltshire, Identification of homogeneous regions for flood frequency analysis, Journal of Hydrology 84(3-4) (1986c) 287-302. https://doi.org/10.1016/0022-1694(86)90128-9.
[49] N.L. Win, K.N. Win, The probability distributions of daily rainfall for Kuantan River Basin in Malaysia, International Journal of Science and Research 3(8) (2014) 977-983.