Pertanika J. Sci. & Technol. 30 (1): 319 - 342 (2022)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

A Comparative Effectiveness of Hierarchical and Non-hierarchical Regionalisation Algorithms in Regionalising the Homogeneous Rainfall Regions

Zun Liang Chuan[1]*, Wan Nur Syahidah Wan Yusoff[1], Azlyna Senawi[1], Mohd Romlay Mohd Akramin[2], Soo-Fen Fam[3], Wendy Ling Shinyie[4] and Tan Lit Ken[5]

[1]Centre for Mathematical Sciences, College of Computing and Applied Sciences, Universiti Malaysia Pahang, Lebuhraya Tun Razak, 26300 Gambang Kuantan, Pahang DM, Malaysia
[2]Faculty of Mechanical and Automotive Engineering Technology, Universiti Malaysia Pahang, 26600 Pekan, Pahang DM, Malaysia
[3]Faculty of Technology Management and Technopreneurship, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Melaka, Malaysia
[4]Department of Mathematics, Faculty of Science, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor DE, Malaysia
[5]Takasago Thermal/Environmental Systems Laboratory, Malaysia-Japan International Institute of Technology (MJIIT), Universiti Teknologi Malaysia, Jalan Sultan Yahya Petra, 54l00 Kuala Lumpur, Malaysia

ABSTRACT

Descriptive data mining has been widely applied in hydrology as the regionalisation algorithms to identify the statistically homogeneous rainfall regions. However, previous studies employed regionalisation algorithms, namely agglomerative hierarchical and non-hierarchical regionalisation algorithms requiring post-processing techniques to validate and interpret the analysis results. The main objective of this study is to investigate the effectiveness of the automated agglomerative hierarchical and non-hierarchical regionalisation algorithms in identifying the homogeneous rainfall regions based on a new statistically significant difference regionalised feature set. To pursue this objective, this study collected 20 historical monthly rainfall time-series data from the rain gauge stations located in the Kuantan district. In practice, these 20 rain gauge stations can be categorised into two statistically homogeneous rainfall regions, namely distinct spatial and temporal variability in the rainfall amounts. The

ARTICLE INFO

Article history:
Received: 06 May 2020
Accepted: 01 October 2020
Published: 04 January 2022

DOI: https://doi.org/10.47836/pjst.30.1.18

E-mail addresses:
chuanzl@ump.edu.my (Zun Liang Chuan)
wnsyahidah@ump.edu.my (Wan Nur Syahidah Wan Yusoff)
azlyna@ump.edu.my (Azlyna Senawi)
akramin@ump.edu.my (Mohd Romlay Mohd Akramin)
famsoofen@utem.edu.my (Soo-Fen Fam)
sy_ling@upm.edu.my (Wendy Ling Shinyie)
tlken@utm.edu.my (Tan Lit Ken)
* Corresponding author

ISSN: 0128-7680
e-ISSN: 2231-8526

results of the analysis show that Forgy *K*-means non-hierarchical (FKNH), Hartigan-Wong *K*-means non-hierarchical (HKNH), and Lloyd *K*-means non-hierarchical (LKNH) regionalisation algorithms are superior to other automated agglomerative hierarchical and non-hierarchical regionalisation algorithms. Furthermore, FKNH, HKNH, and LKNH yielded the highest regionalisation accuracy compared to other automated agglomerative hierarchical and non-hierarchical regionalisation algorithms. Based on the regionalisation results yielded in this study, the reliability and accuracy that assessed the risk of extreme hydro-meteorological events for the Kuantan district can be improved. In particular, the regional quantile estimates can provide a more accurate estimation compared to at-site quantile estimates using an appropriate statistical distribution.

Keywords: Anderson Darling statistical test, bootstrap, hierarchical, non-hierarchical, regionalisation algorithm, unbiased statistical test

INTRODUCTION

The rapid advancement of technology and data collection has facilitated organisations and researchers to gather huge amounts of data. However, it is extremely challenging to extract fruitful and reliable insights using conventional statistical analysis techniques (Tan et al., 2006). Therefore, data mining was rapidly growing in performing descriptive and predictive tasks. In principle, descriptive data mining derives patterns, such as associations, trajectories, trends, clusters, and anomalies that summarise principal relationships in the data sets. In nature, descriptive data mining requires post-processing techniques to validate and interpret the results of the analysis. Meanwhile, predictive data mining is used to predict future values of particular attributes in the presence of uncertainty based on historical attributes.

In previous hydrology studies, descriptive data mining tools including agglomerative hierarchical and non-hierarchical regionalisation algorithms have been widely used in identifying homogeneous rainfall regions (Ahmad et al., 2013; Awan et al., 2014; Burn et al., 1997; Chuan et al., 2018a; Chuan et al., 2018b; Guttman, 1993; Hamdan et al., 2015; Ngongondo et al., 2011; Nnaji et al., 2014; Terassi & Galvani, 2017). The principal objective of using descriptive data mining tools in hydrology studies is to extrapolate insights from the gauge into ungauged rainfall stations based on the limited amount of historical rainfall time series data, which can increase the reliability of the risk assessment of extreme hydro-meteorological events. For instance, Guttman (1993) proposed the use of average linkage agglomerative hierarchical (ALAH) and Ward's minimum variance agglomerative hierarchical (WMVAH) regionalisation algorithms to identify the homogeneous precipitation regions in the United States. To identify the homogeneous precipitation regions, they used the Euclidean distance as the dissimilarity measure.

Variables, such as the geographical insights of precipitation gauge stations, mean annual precipitation amount, and average variability of the annual precipitation cycle, were regarded as regionalisation features. In addition, Burn et al. (1997) proposed the use of the agglomerative hierarchical regionalisation algorithm in identifying the homogeneous watersheds in Canada based on Webster and Burrough's distance as the dissimilarity measure. In addition, seasonality measures of the catchments were regarded as regionalisation features. Both previous studies used the L-moments based homogeneity measure to validate the regionalised homogeneous regions. Moreover, Nnaji et al. (2014) proposed the use of the ALAH regionalisation algorithm to identify the homogeneous regions in the Federal Republic of Nigeria based on the coefficient of variation extracted from the historical monthly rainfall time series data and Euclidean distance dissimilarity measure. However, this article does not describe the homogeneity validation of the identified homogeneous rainfall regions. Recently, Terassi and Galvani (2017) proposed identifying the homogeneous rainfall regions in the watersheds of the eastern region of the state of Paraná using the WMVAH regionalisation algorithm. In identifying the homogeneous rainfall regions, they employed the Euclidean distance dissimilarity measure and regionalisation features, such as rainfall variability, relief characteristics, the spatial proximity of pluviometric and meteorological stations.

Ngongondo et al. (2011) proposed the use of WMVAH and MacQueen *K*-means non-hierarchical (MKNH) regionalisation algorithms in identifying the homogeneous rainfall regions in Southern Malawi. First, they define the homogeneous rainfall regions based on the regionalisation features, such as geographical insights of the rain gauge stations and at-site mean annual precipitation. Then, the number of optimum homogeneous rainfall regions is determined using Hubert's gamma coefficient and Dunn internal clustering validation indices. Moreover, they validate the homogeneity of the identified rainfall regions using discordant and L-moments-based homogeneity measures. Awan et al. (2014) conducted a comparison study of the effectiveness between agglomerative hierarchical and non-hierarchical regionalisation algorithms. In particular, they investigated the effectiveness of WMVAH and the MKNH regionalisation algorithms to identify the homogeneous rainfall regions in the East Asia monsoon, including China, Korea, Japan, and Taiwan using the historical monthly rainfall time series data based on the Euclidean distance dissimilarity measure and average annual rainfall amounts. This study employed Calinski–Harabasz, Krzanowski–Lai, and Davies–Bouldin internal clustering validation indices to determine the optimum number of homogeneous regions. The results for the analysis of this study show that the MKNH regionalisation algorithm is superior to the WMVAH regionalisation algorithm.

In Malaysia, previous studies have been extensively employing agglomerative hierarchical regionalisation algorithms in identifying homogeneous rainfall regions. In particular, Ahmad et al. (2013) proposed identifying the homogeneity of historical annual rainfall time series data recorded from 59 rain gauge stations in Peninsular Malaysia using a complete linkage agglomerative hierarchical (CLAH) regionalisation algorithm and correlation coefficient dissimilarity measure. Their study selected and evaluated a superior regionalisation algorithm from 77 potential agglomerative hierarchical regionalisation algorithms on Malaysia geographical, monsoon limitation, and segregation of stations. They proposed selecting the optimum number of regions using eight internal clustering validation indices. In addition, Hamdan et al. (2015) proposed employing the CLAH regionalisation algorithm to identify homogeneous rainfall regions using the historical annual time series data recorded from 75 rain gauge stations located in Peninsular Malaysia and Euclidean distance dissimilarity measure. However, Ahmad et al. (2013) and Hamdan et al. (2015) did not present the homogeneity validation of the identified homogeneous rainfall regions.

Chuan et al. (2018b) proposed the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) multi-criteria decision-making algorithm and internal clustering validation indices, such as C-index, Davies-Bouldin index, Dunn-index, and Gamma index to determine the superior agglomerative hierarchical regionalisation algorithm and the optimum number of homogeneous rainfall regions. Their findings show that the ALAH regionalisation algorithm, which is based on the Altgower dissimilarity measure, had successfully delineated 20 historical monthly rainfall time series data recorded from the rain gauge stations located in the Kuantan district into two coastal and inland regions.

Furthermore, it validated the identified homogeneous regions using a non-parametric Bootstrap-based K-sample Anderson Darling (BKAD) statistical test. Contrariwise, Chuan et al. (2018a) proposed an automated regionalisation algorithm, which integrated the ALAH regionalisation algorithm and multi-scale bootstrap resampling to identify statistically homogeneous rainfall regions. Their proposed automated regionalisation algorithm had regionalised 20 historical monthly rainfall time series data of the Kuantan district in the coastal and inland regions. These two regions show different spatial and temporal rainfall variability. In general, the automated regionalisation algorithm is more advantageous compared to other previously proposed algorithms. The automated regionalisation algorithm is competent to identify and validate the homogeneous rainfall regions and vice versa for the non-automated regionalisation algorithms. In addition, the automated regionalisation algorithm uses an approximately unbiased (AU) statistical test that provides statistical evidence and vice versa for the L-moments based homogeneity measure.

The main objective of this study is to investigate the effectiveness of the automated agglomerative hierarchical and non-hierarchical regionalisation algorithms in identifying the homogeneous rainfall regions based on a new statistically significant difference regionalised feature set. In specific, the agglomerative hierarchical regionalisation algorithms employed in this study include average linkage agglomerative hierarchical (ALAH), complete linkage agglomerative hierarchical (CLAH), single linkage agglomerative hierarchical (SLAH), and Ward's minimum variance agglomerative hierarchical (WMVAH). The non-hierarchical regionalisation algorithms include Forgy *K*-means (FKNH), Hartigen-Wong *K*-means (HKNH), Lloyd *K*-means (LKNH), and MacQueen *K*-means (MKNH). Moreover, the centred and uncentred inverse correlation coefficients are used as the dissimilarity measure for agglomerative hierarchical algorithms, and Euclidean distance is used as the dissimilarity measure for non-hierarchical algorithms. This study regionalised the homogeneous rainfall regions based on the rainfall distribution characteristics of the historical monthly rainfall time series data. Therefore, the superior regionalisation algorithm is determined based on the accuracy rate of regionalisation. To pursue the main objective of this study, the rest of this paper is organised as follows. Section 2 presents the detailed insights of the rain gauge stations located in the Kuantan district, while Section 3 provides a brief overview of the theoretical background. The results of the analysis and discussion are presented in Section 4. Finally, the concluding remarks are presented in Section 5.

STUDY AREAS

The Kuantan river basin is about 440 km long and irrigates 29,300 km^2 of the area, and it is located in the eastern part of Peninsular Malaysia. This river is in the Kuantan district that covers 1,630 km^2 of the catchment area starting from the forest reserved area in Mukim Ulu Kuantan; it passes through the agricultural area and the state capital of Pahang before it is discharged into the South China Sea (Saeed et al., 2016; Chuan et al., 2020). Moreover, this river is the main principle tributary, which irrigates the major rural, urban, agricultural, and industrial areas of the Kuantan district. The Kuantan district has a tropical rainforest climate belonging to the Köppen climate classification, which experiences two seasons per year, namely dry and hot seasons during the Southwest Monsoon and rainy seasons during the Northeast Monsoon. However, the anthropogenic emissions of greenhouse gasses are caused by human activities, which are the main contributor to global warming. Consequently, global warming has been increasing the risk of extreme hydro-meteorological events. Therefore, there is a need for a reliable and accurate risk assessment for extreme hydro-meteorological events using sufficient historical rainfall time series data.

This study used 20 historical monthly rainfall time series data (in mm) with 58 months in the Kuantan district to evaluate the regionalisation algorithms. In particular, the historical monthly rainfall time series data cover the period from February 2010 to November 2014, which was collected from the Department of Irrigation and Drainage Malaysia. Figure 1 shows the locations of 20 rain gauge stations in this study. Meanwhile, Table 1 presents the geographical insights of the 20 rain gauge stations and the characteristics of rainfall distribution based on the historical monthly rainfall time series data.

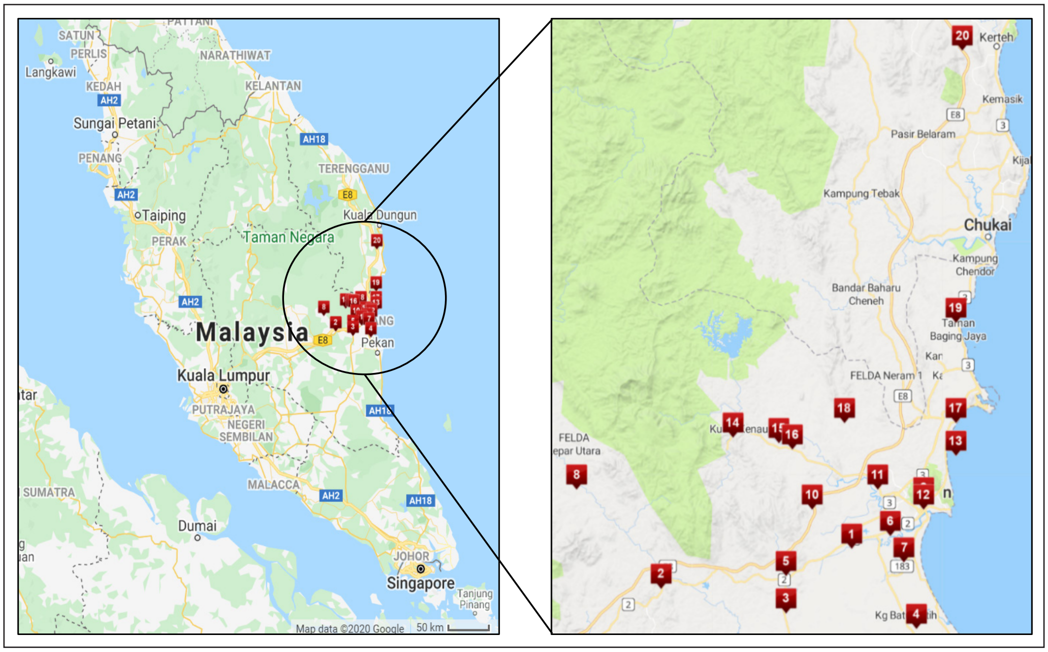


Figure 1. Locations of 20 rain gauge stations located in the Kuantan River Basin for Kuantan District

Table 1

The geographical coordinates insights of the 20 rain gauge stations and the characteristics of the probability distribution of the historical monthly rainfall time series data

Station	Station ID	Station Name	Regionalisation Features						
			θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7
1	3432001	Sri Damai	3.75	103.22	14.90	90.56	63.94	33.03	16.12
2	3629098	Paya Bungor	3.69	102.93	34.70	158.11	66.46	11.61	10.94
3	3631001	Kg. Pulau Manis	3.65	103.12	37.40	181.52	65.49	13.93	16.31
4	3633104	Kg. Bahru, Penor	3.63	103.32	7.60	179.62	78.90	28.75	24.50
5	3731018	JKR Gambang	3.71	103.12	41.30	234.01	80.29	18.44	19.46
6	3732020	Paya Besar di Kuantan	3.77	103.28	6.00	162.88	76.36	19.23	13.93
7	3732021	Kg. Sg. Soi	3.73	103.30	11.90	210.96	90.27	33.53	28.89
8	3828091	Ldg. Ulu Lepar	3.84	102.80	91.70	167.20	60.28	10.71	10.04
9	3830001	Ldg. Mentiga	3.82	103.33	9.40	199.19	79.08	12.46	13.56
10	3831002	Felda Pancing	3.81	103.16	71.40	234.07	103.13	26.00	19.28
11	3832015	Rancangan Pam Paya Pinang	3.84	103.26	6.70	209.73	90.43	28.47	25.47
12	3833002	Pejabat JPS Negeri Pahang	3.81	103.33	10.30	180.86	83.88	23.27	21.47
13	3833004	Ldg. Jeram di Kuantan	3.89	103.38	-1.40	210.94	109.27	43.74	35.32

Table 1 (*continue*)

Station	Station ID	Station Name	Regionalisation Features						
			θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7
14	3930012	Sg. Lembing P.C.C.L Mill	3.92	103.04	33.10	245.67	87.26	7.70	9.68
15	3931013	Ldg. Nada	3.91	103.11	16.90	227.94	87.58	18.06	18.78
16	3931014	Ldg. Kuala Reman	3.90	103.13	29.90	201.86	83.56	9.32	9.73
17	3933003	Balok di Kuantan	3.94	103.38	4.10	220.82	112.23	47.76	33.70
18	4031001	Bkt. Sagu	3.94	103.21	20.90	511.75	193.46	42.31	51.69
19	4033001	Kg. Cherating	4.09	103.38	9.00	221.22	111.61	46.15	30.14
20	4033002	Kg. Sg. Ular	4.50	103.39	58.50	228.74	110.54	45.15	34.24

Note: θ_1 = latitude; θ_2 = longitude; θ_3 = altitude; θ_4 = median; θ_5 = coefficient of variation; θ_6 = skewness; θ_7 = kurtosis

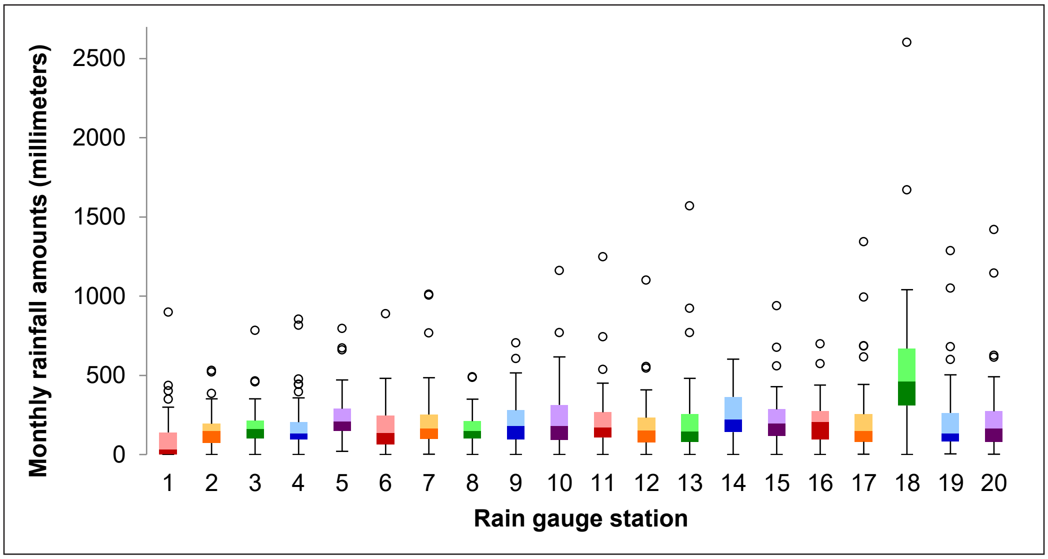


Figure 2. The graphical descriptive summary of twenty historical monthly rainfall amounts recorded from rain gauge stations located Kuantan District

Based on the statistics, the median is more robust than the skewed data and outliers compared to the arithmetic mean (Saeed et al., 2016). Figure 2 and Table 1 show that the employed historical monthly rainfall time series data are highly skewed to the right, and there are outliers in all historical monthly rainfall time series data. In particular, Figure 2 illustrates that the upper box is wider than the lower box for the 20 historical rainfall time series data, while θ_6 indicates the positive values of skewness. In addition, Figure 2 also presents the presence of outliers (unfilled circle) that excluded the historical monthly rainfall time series data recorded from the 14th rain gauge station. Therefore, the most appropriate central value is the median rather than the arithmetic mean.

METHODOLOGY

This section provides a brief overview of the theoretical background of the agglomerative hierarchical and non-hierarchical regionalisation algorithm employed in this study. In addition, it also presents the theoretical background of the data screening and statistical homogeneity tests, such as approximately unbiased (AU) and non-parametric Bootstrap-based K-sample Anderson Darling (BKAD) statistical test. This study used the AU statistical test to integrate agglomerative hierarchical clustering analysis to establish the automated regionalisation algorithms. The dissimilarity measures include the centred and uncentred inverse correlation coefficient. However, the dissimilarity measure applied in the non-hierarchical regionalisation algorithms is the Euclidean distance. The mechanism of the *K*-means non-hierarchical regionalisation algorithm defines the homogeneous rainfall regions by minimising the sum of squared error (*SSE*) based on the Euclidean space. Therefore, the Euclidean distance is appropriately applied in this algorithm.

Data Screening

Data screening is a preliminary and essential procedure in inspecting and treating errors in data set, such as missing data. In practice, the missing data in the historical monthly rainfall time series data is invertible due to the meteorological extremes, recorded error, and malfunctions of instruments (Chuan et al., 2020; Chuan et al., 2018a; Chuan et al., 2018b; Chuan et al., 2018c; Saeed et al., 2016). $\mathbf{X}^{incomp}$ is a matrix of size $I \times J$ of the historical monthly rainfall amount as in Equation 1:

$$\mathbf{X}^{incomp} = \left[\left(1-\pi\right) X_{ij}^{obs} + \pi X_{ij}^{miss} \right]_{I \times J} \tag{1}$$

where, X_{ij}^{obs} and X_{ij}^{miss}, respectively represent the observed and missing historical monthly rainfall amount for i th month recorded from j th rain gauge stations. $\pi \leq 0.3$ represents the rate of missingness and $i, \left(j\right) = 1, 2, \ldots, I, \left(J\right)$. To obtain a complete data set, this study employed the column median single imputation algorithm (Saeed et al., 2016) in treating the missing data for the matrix of $\mathbf{X}^{incomp}$ as in Equation 2:

$$X_{j}^{miss} = \underset{1 \leq i \leq I}{median} \left[X_{ij}^{obs} \right] \tag{2}$$

Therefore, this study obtained a complete matrix for the historical monthly rainfall amount, $\mathbf{X}^{comp}$. This study employed a column median single imputation algorithm in treating the missing data rather than multiple imputation algorithms. The column median single imputation algorithm requires a low computational cost compared to multiple imputation algorithms. Similar regionalisation results are obtained after treating the missing data using column median single and multiple imputation algorithms.

Regionalisation Algorithms

Descriptive data mining is used to derive patterns, such as homogeneous rainfall regions that summarise the underlying relationships in $\mathbf{X}^{comp}$, which the adequacy of regionalisation of homogeneous rainfall regions is highly dependent on the features extracted from $\mathbf{X}^{comp}$. The extraction of the insignificant features from $\mathbf{X}^{comp}$ can degrade the adequacy of regionalisation results (Dash & Liu, 2003). The regionalisation features set, $\boldsymbol{\theta}$, employed in this study include θ_1, θ_2, θ_3, θ_4, θ_5, θ_6 and θ_7, which indicate the statistically significant difference results (F test-statistic = 4,197,057.90; p-value = 0.000) after testing using Roy's greatest root of multivariate analysis of variance (Chuan et al., 2018c; Chuan et al., 2020). Therefore, the homogeneous rainfall regions are identified based on $\boldsymbol{\theta}$ using the agglomerative hierarchical and non-hierarchical regionalisation algorithms.

Agglomerative Hierarchical Regionalisation Algorithms. Previous hydrology studies stated that agglomerative hierarchical regionalisation algorithms are widely employed to define the homogeneous rainfall regions (Ahmad et al., 2013; Burn et al., 1997; Chuan et al., 2018a; Chuan et al., 2018b; Guttman, 1993; Hamdan et al., 2015; Nnaji et al., 2014; Terassi & Galvani, 2017). In principle, agglomerative hierarchical regionalisation algorithms perform $J-1$ successive fusion based on predetermined dissimilarity measures to establish a single agglomerative hierarchical dendrogram. Assume that $\min\limits_{1\le j\le J}\left[\lambda\left(\boldsymbol{\theta}_{j_{m0}},\boldsymbol{\theta}_{j_{n0}}\right)\right]$ is the minimum distance based on the predetermined dissimilarity measure, λ, for a single cluster, $\boldsymbol{\theta}_{j_m}$, which comprises the fused pair of clusters, $\boldsymbol{\theta}_{j_{m0}}$ and $\boldsymbol{\theta}_{j_{n0}}$. By updating ALAH, CLAH, SLAH, and WMVAH regionalisation algorithm mechanism functions in Equations 3 to 6 (Chuan et al., 2018b), it will produce a new hierarchical agglomerative dendrogram, $\lambda^*\left(\boldsymbol{\theta}_{j_m},\ \boldsymbol{\theta}_{j_n}\right)$, between cluster $\boldsymbol{\theta}_{j_m}$ and remaining infused clusters, $\boldsymbol{\theta}_{j_n}$.

$$\lambda^*_{\text{ALAH}}\left(\boldsymbol{\theta}_{j_m},\ \boldsymbol{\theta}_{j_n}\right)=\frac{a_{\boldsymbol{\theta}_{j_{m0}}}\lambda\left(\boldsymbol{\theta}_{j_{m0}},\ \boldsymbol{\theta}_{j_m}\right)\ +\ a_{\boldsymbol{\theta}_{j_{n0}}}\lambda\left(\boldsymbol{\theta}_{j_{n0}},\boldsymbol{\theta}_{j_n}\right)}{a_{\boldsymbol{\theta}_{j_{m0}}}+a_{\boldsymbol{\theta}_{j_{n0}}}} \tag{3}$$

$$\lambda^*_{\text{CLAH}}\left(\boldsymbol{\theta}_{j_m},\boldsymbol{\theta}_{j_n}\right)=\frac{\lambda\left(\boldsymbol{\theta}_{j_{m0}},\boldsymbol{\theta}_{j_m}\right)+\lambda\left(\boldsymbol{\theta}_{j_{n0}},\boldsymbol{\theta}_{j_n}\right)+\left|\lambda\left(\boldsymbol{\theta}_{j_{m0}},\boldsymbol{\theta}_{j_m}\right)-\lambda\left(\boldsymbol{\theta}_{j_{n0}},\boldsymbol{\theta}_{j_n}\right)\right|}{2} \tag{4}$$

$$\lambda^*_{\text{SLAH}}\left(\boldsymbol{\theta}_{j_m},\boldsymbol{\theta}_{j_n}\right)=\frac{\lambda\left(\boldsymbol{\theta}_{j_{m0}},\boldsymbol{\theta}_{j_m}\right)+\lambda\left(\boldsymbol{\theta}_{j_{n0}},\boldsymbol{\theta}_{j_n}\right)-\left|\lambda\left(\boldsymbol{\theta}_{j_{m0}},\boldsymbol{\theta}_{j_m}\right)-\lambda\left(\boldsymbol{\theta}_{j_{n0}},\boldsymbol{\theta}_{j_n}\right)\right|}{2} \tag{5}$$

$$\lambda^*_{\text{WMVAH}}\left(\boldsymbol{\theta}_{j_m},\boldsymbol{\theta}_{j_n}\right)=\frac{\lambda\left(\boldsymbol{\theta}_{j_{m0}},\boldsymbol{\theta}_{j_m}\right)\left(a_{\boldsymbol{\theta}_{j_{m0}}}+a_{\boldsymbol{\theta}_{j_n}}\right)+\ \lambda\left(\boldsymbol{\theta}_{j_{n0}},\boldsymbol{\theta}_{j_n}\right)\left(a_{\boldsymbol{\theta}_{j_{n0}}}+a_{\boldsymbol{\theta}_{j_n}}\right)-\lambda\left(\boldsymbol{\theta}_{j_{m0}},\boldsymbol{\theta}_{j_{n0}}\right)\left(a_{\boldsymbol{\theta}_{j_n}}\right)}{a_{\boldsymbol{\theta}_{j_{m0}}}+a_{\boldsymbol{\theta}_{j_{n0}}}+a_{\boldsymbol{\theta}_{j_n}}} \tag{6}$$

where, $a_{\boldsymbol{\theta}_{j_{m0}}}$, $a_{\boldsymbol{\theta}_{j_{n0}}}$ and $a_{\boldsymbol{\theta}_{j_n}}$ represent the number of $\boldsymbol{\theta}$ extracted from the historical monthly rainfall time series data recorded from the $j\,\text{th}$ rain gauge that is included in the clusters of $\boldsymbol{\theta}_{j_{m0}}$, $\boldsymbol{\theta}_{j_{n0}}$ and $\boldsymbol{\theta}_{j_n}$, respectively. Moreover, there are two differences of λ employed in this study to investigate their effectiveness in regionalising the homogeneous rainfall regions. They include centred, λ_1, and uncentred, λ_2, inverse correlation coefficients as in Equations 7 and 8:

$$\lambda_1 = 1 - \frac{\sum\left(\text{vec}\left(\boldsymbol{\theta}_{j_1}\right) - \overline{\text{vec}\left(\boldsymbol{\theta}_{j_1}\right)}\right)\left(\text{vec}\left(\boldsymbol{\theta}_{j_2}\right) - \overline{\text{vec}\left(\boldsymbol{\theta}_{j_2}\right)}\right)}{\sqrt{\sum\left(\text{vec}\left(\boldsymbol{\theta}_{j_1}\right) - \overline{\text{vec}\left(\boldsymbol{\theta}_{j_1}\right)}\right)^2 \sum\left(\text{vec}\left(\boldsymbol{\theta}_{j_2}\right) - \overline{\text{vec}\left(\boldsymbol{\theta}_{j_2}\right)}\right)^2}} \tag{7}$$

$$\lambda_2 = 1 - \frac{\sum\left(\text{vec}\left(\boldsymbol{\theta}_{j_1}\right)\text{vec}\left(\boldsymbol{\theta}_{j_2}\right)\right)}{\sqrt{\sum\left(\text{vec}\left(\boldsymbol{\theta}_{j_1}\right)\right)^2 \sum\left(\text{vec}\left(\boldsymbol{\theta}_{j_2}\right)\right)^2}} \tag{8}$$

where, $\text{vec}(\cdot)$ represents the vectorisation function and $j_1 \neq j_2$.

Non-hierarchical Regionalisation Algorithms. The non-hierarchical regionalisation algorithms, such as *K*-means is an iterative regionalisation algorithm with the main principle to regionalise J rain gauge stations into *K* predefined distinct non-overlapping homogeneous rainfall regions, $\mathbf{C}_k;\ k = 1, 2, \ldots, K$. Based on the mechanism of *K*-means non-hierarchical regionalisation algorithm, J rain gauge stations are delineated into *K* homogeneous rainfall regions by minimising the sum of squared error (*SSE*) as in Equation 9:

$$SSE = \sum_{k=1}^{K}\sum_{j=1}^{J}\left\|\boldsymbol{\theta}_j^k - \omega_k\right\| \tag{9}$$

where, $\left\|\boldsymbol{\theta}_j^k - \omega_k\right\|$ represents the Euclidean distance between the infused rain gauge stations and the centre of homogenous rainfall regions, ω. On the other hand, *K*-means non-hierarchical regionalisation algorithms require prior knowledge in predetermining the *K* numbers of $\mathbf{C}_k$. This study used four distinct non-hierarchical regionalisation algorithms in performing the *K*-means process: FKNH (Forgy, 1965), MKNH (MacQueen, 1967), HKNH (Hartigan & Wong, 1979), and LKNH (Lloyd, 1982). The identified homogeneous rainfall regions are based on the regionalisation algorithms validated using BKAD statistical test (Chuan et al., 2018b), which provides statistical evidence.

Homogeneity Validation Statistical Test

In principle, descriptive data mining requires post-processing techniques in validating and interpreting the results of the analysis. Therefore, this study employed AU and BKAD statistical tests to validate the homogeneity of the identified homogeneous rainfall regions. In particular, the AU statistical test is associated with agglomerative hierarchical regionalisation algorithms to determine the optimum number of homogeneous rainfall regions and validate the homogeneity of identifying rainfall regions. Meanwhile, BKAD statistical test was employed to validate the homogeneity of identifying rainfall regions. Furthermore, AU statistical test is deemed appropriate for the phylogenetic tree selection (Shimodaira, 2002); therefore, two distinct homogeneity tests were used to pursue the main objective of this study.

Approximately Unbiased Statistical Test. AU statistical test was introduced by Shimodaira (2002), who employed the multi-scale bootstrap technique to reduce the bias in a statistical test. This statistical test is applied to the maximum likelihood tree selection in acquiring the confidence set of the trees (Shimodaira, 2002). In general, AU statistical test is more advantageous than the L-moments based homogeneity measure to validate the homogeneity of identifying rainfall regions. AU statistical test can provide statistical evidence and vice versa for L-moments based homogeneity measures. The following is the procedure in determining the optimum number of homogeneous rainfall regions and the validation of identifying homogeneity based on the AU statistical test.

Step 1: Fix the constants of scaling, $\Psi_r; r = 1, 2, \ldots, R$, and the number of replicates for the multi-scale bootstrap resampling, Ω_r, where $\Omega_r = 10,000$ is used in this study.
Step 2: Generate Ω_r bootstrap that replicates with a sequence length of $\psi^* = \Psi_r.J$, which is denoted as $\phi^s\left(\Psi_r\right)$ for $\forall r$ and $s = \Omega_r$.
Step 3: Calculate the bootstrap probability, $\Pr_\Omega\left(\Psi_r\right)$, based on Equation 10.

$$\Pr_\Omega\left(\Psi_r\right) = \frac{\#\left\langle \phi^s\left(\Psi_r\right) \in H_1 \right\rangle}{\Omega_r}; H_1 = \sqrt{\sum_{d=1}^{D} \mu_d} \leq \Lambda_D \tag{10}$$

where, Λ_D represents the region in D-dimensional parameter spaces.
Step 4: Estimate the curvature, ξ_1, and the signed distance, ξ_2, by minimising the sum squared of residual (SSR) as in Equation 11:

$$SSR\left(\xi_1, \xi_2\right) = \sum_{g=1}^{G} \left(\frac{\frac{\xi_1}{\sqrt{\Psi_r}} + \xi_2 - \Phi^{-1}\left(1 - \Pr_\Omega\left(\Psi_r\right)\right)}{\hat{\sigma}_g} \right)^2 \tag{11}$$

where, $\hat{\sigma}_g = \frac{\Omega_r \varphi\left(\Phi^{-1}\Pr_\Omega\left(\Psi_r\right)\right)}{\sqrt{\Pr_\Omega\left(\Psi_r\right)\left(1-\Pr_\Omega\left(\Psi_r\right)\right)}}$. Meanwhile, $\varphi(\cdot)$ and $\Phi^{-1}(\cdot)$ represent the density and quantile function of the standard normal distribution, respectively.

Step 5: Calculate the p-value of the AU statistical test based on Equation 12.

$$\Pr_{\text{AU}} = 1 - \Phi\left(\xi_2 - \xi_1\right) \tag{12}$$

K-sample Anderson Darling Statistical Test. The BKAD statistical test is the generalisation of the classical *K*-sample Anderson Darling statistical test, and this statistical test is free from any statistical assumption (Chuan et al., 2018b). For example, suppose that $x_{(1)}^{comp} < x_{(2)}^{comp} < \ldots < x_{(IK)}^{comp}$ is the pooled order of sample $\mathbf{C}_k$. Therefore, the BKAD statistical test, Q_{BKAD}, and its standard deviation, $Sd\left(Q_{\text{BKAD}}\right)$, can be defined as Equations 13 and 14:

$$Q_{\text{BKAD}} = \sum_{i=1}^{IK-1} \frac{\left(K\rho_{ik} - i\right)^2}{i\left(IK - i\right)} \tag{13}$$

$$Sd\left(Q_{\text{BKAD}}\right) = \sqrt{\frac{\Gamma\left(IK-3\right)\left(\upsilon_1 + \upsilon_2 IK + \upsilon_3\left(IK\right)^2 + \upsilon_4\left(IK\right)^3\right)}{\Gamma\left(IK\right)}} \tag{14}$$

where, ρ_{ik} represents the number of historical monthly rainfall amounts in the k th sample that is not more than the *j*th smallest historical monthly rainfall amount in the pooled sample. Meanwhile, υ_1, υ_2, υ_3 and υ_4 are expressed as Equations 15 to 18:

$$\upsilon_1 = 2K\left(3K + \left(K-2\right)\sum_{i=1}^{IK-1}\frac{1}{i}\right) \tag{15}$$

$$\upsilon_2 = -\frac{2}{I}\left(IK^2 - 3IK + 3K - \left(3IK^2 + 2IK + 2I + K\right)\sum_{i=1}^{IK-1}\frac{1}{i} - IK\sum_{i=1}^{IK-1}\frac{1}{\left(IK-i\right)\left(i+1\right)}\right) \tag{16}$$

$$\upsilon_3 = -\frac{2}{I}\left(2IK^2 + 3I + 3K - \left(4IK - 4I - 7K\right)\sum_{i=1}^{IK-1}\frac{1}{i} - \left(IK^2 + K + 2I\right)\sum_{i=1}^{IK-1}\frac{1}{\left(IK-i\right)\left(i+1\right)}\right) \tag{17}$$

$$\upsilon_4 = -\frac{2}{I}\left(3IK + 3I + 5K - \left(2IK - 2I - 3K\right)\sum_{i=1}^{IK-1}\frac{1}{\left(IK-i\right)\left(i+1\right)}\right) \tag{18}$$

In principle, the identified rainfall regions are statistically homogeneous if and only if (Equation 19):

$$\eta_{\text{BKAD}} = \frac{Q_{\text{BKAD}}^2 - K + 1}{Sd^2\left(Q_{\text{BKAD}}\right)} \geq \eta_{K-1,\alpha} \quad (19)$$

where, the upper tail percentage points $\eta_{K-1,\ \alpha} = \dfrac{Q_{K-1}^2 - (K-1)}{\sqrt{\dfrac{1.7392(K-1)}{3}}}$, and Q_{K-1}^2 is acquired from fitting Pearson curves (Scholz & Stephens, 1987). Based on Equation (19), the stability property of the BKAD statistical test is questionable. Therefore, the BKAD statistical test was performed based on the ranks of sample historical monthly rainfall amount. In addition, this study employed the non-parametric bootstrap resampling with 10,000 replications in determining the acceptability limit of the BKAD test to overcome this limitation.

RESULTS AND DISCUSSION

The analysis of this study was mainly conducted using R statistical software. Figure 2 presents the five-number summary of the 20 historical monthly rainfall time series data recorded from the rain gauge stations located in the Kuantan district that includes minimum, first quartile, median, third quartile, and maximum amount of the monthly rainfall. Based on Figure 2, the historical monthly rainfall time series data recorded from the first and 18th rain gauge stations show the least and highest rainfall amounts on average, respectively. Meanwhile, the historical monthly rainfall time series data recorded from the fourth and 18th rain gauge stations show the least and highest variability of rainfall amounts, respectively. According to Chuan et al. (2018a), the 20 rain gauge stations can be regionalised into two statistically homogeneous rainfall regions, namely the coastal (second, third, fifth, eighth, 10th, 14th, 15th, and 16th) and inland regions (first, fourth, sixth, seventh, ninth, 11th, 12th, 13th, 17th, 18th, 19th, and 20th) as illustrated in Figure 3. In particular, these coastal and inland regions comprise distinct spatial and temporal variability (Chuan et al., 2018b).

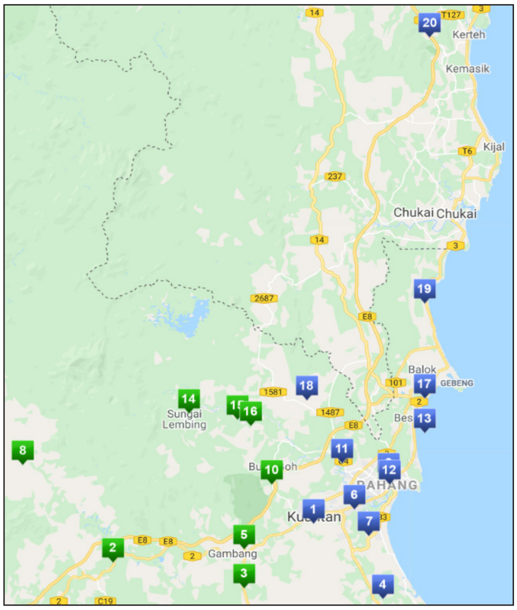
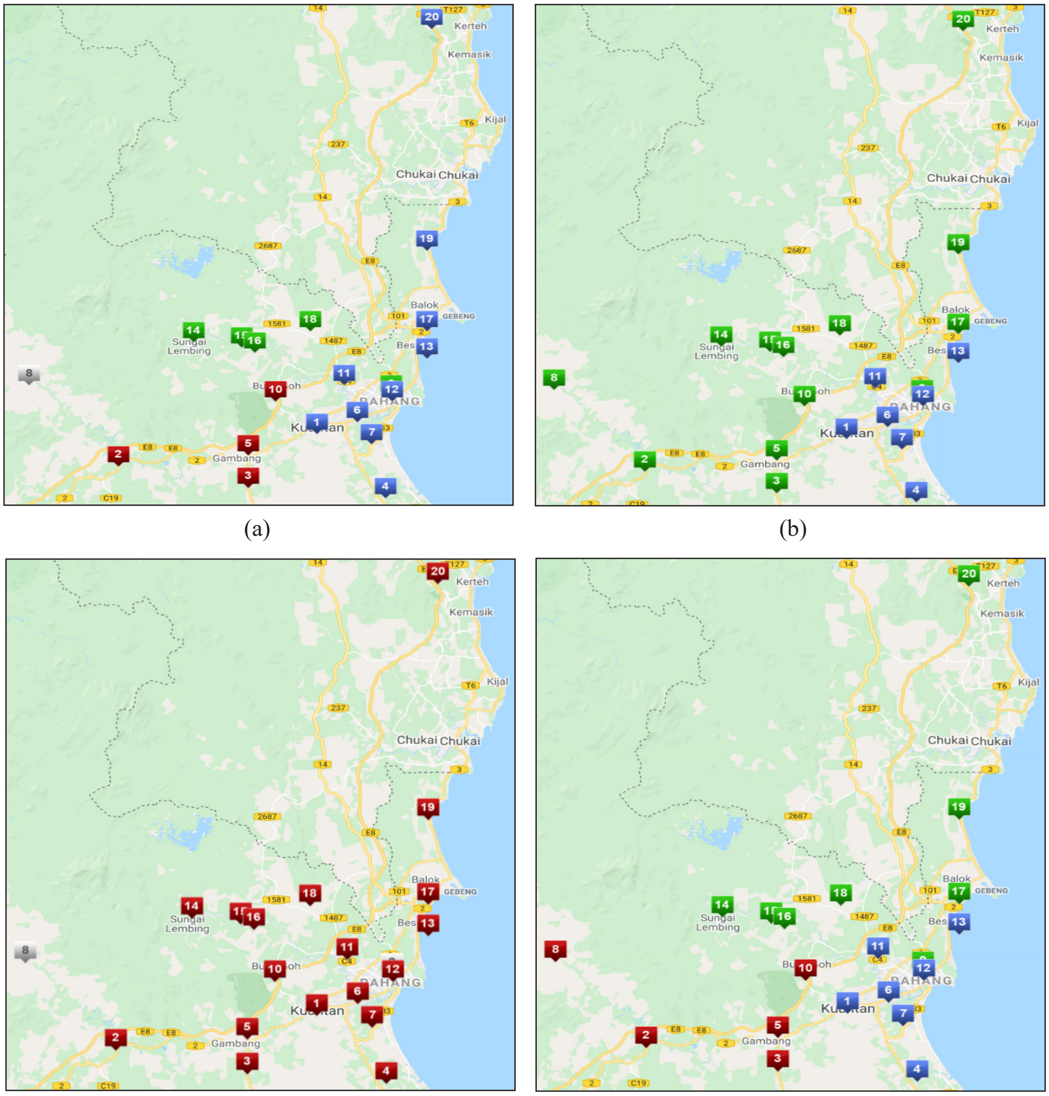


Figure 3. The best regionalised homogeneous rainfall regions, which comprises the coastal and inland regions

Figures 4 and 5 presented the regionalisation analysis results from the dendrograms (Figures A1 & A2, Appendix) based on ALAH, CLAH, SLAH, and WMVAH regionalisation algorithms using



(a)

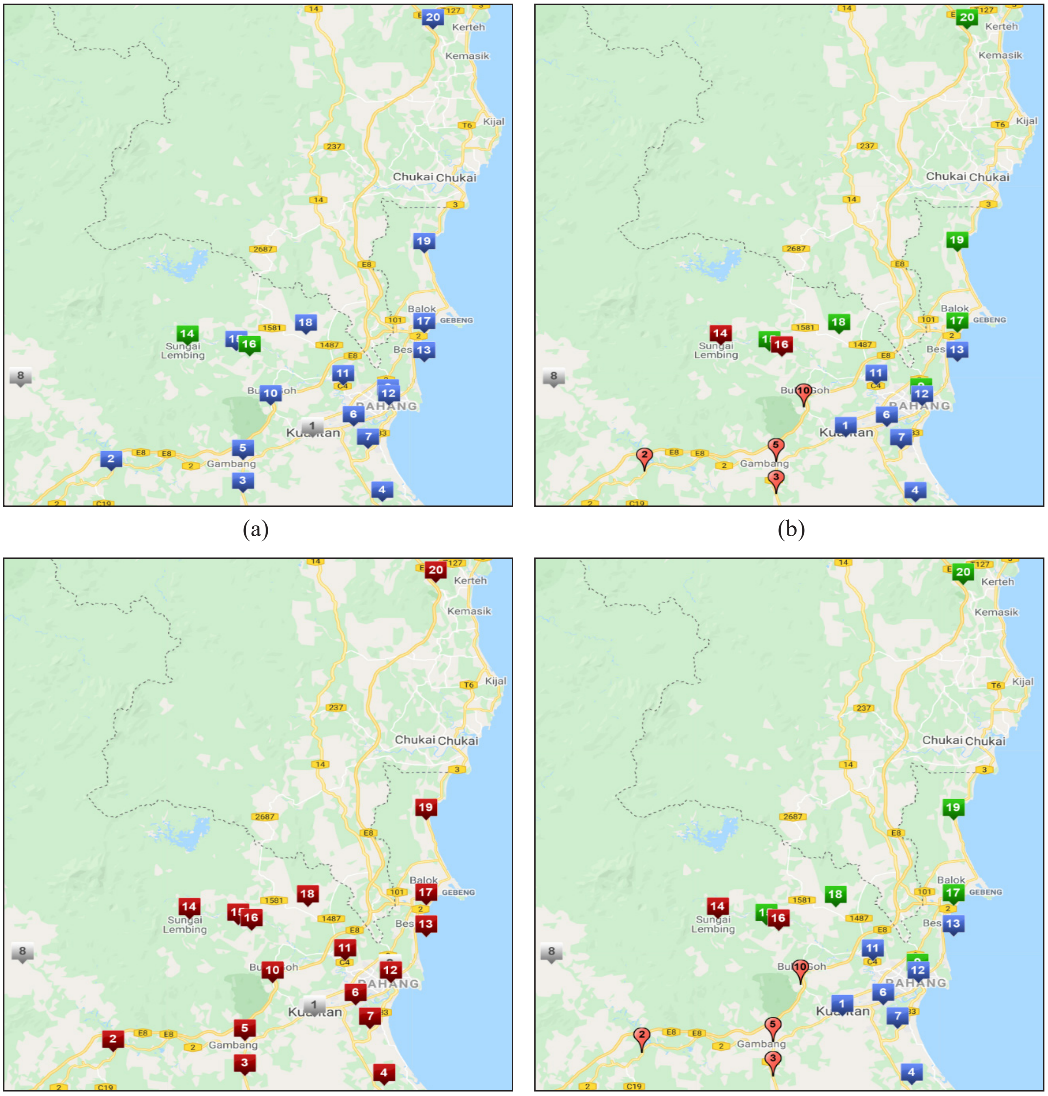
(b)

(c)

(d)

Figure 4. The identified regionalised homogeneous rainfall regions using (a) ALAH, (b) CLAH, (c) SLAH, (d) WMVAH regionalisation algorithm based on λ_1

λ_1 and λ_2, respectively. Meanwhile, Figure 6 presents the regionalisation analysis results based on FKNH, HKNH, LKNH, and MKNH regionalisation algorithms. Based on Figures 4(a) and 4(c), the white marker of the eighth rain gauge station shows that this station is unfit to merge with other statistically homogeneous rainfall regions. A similar representative is also presented in Figures 5(a), 5(b), 5(c), and 5(d). Specifically, the rain gauge stations are unfit to merge with other identified statistically homogeneous rainfall regions represented using white markers.



(a)

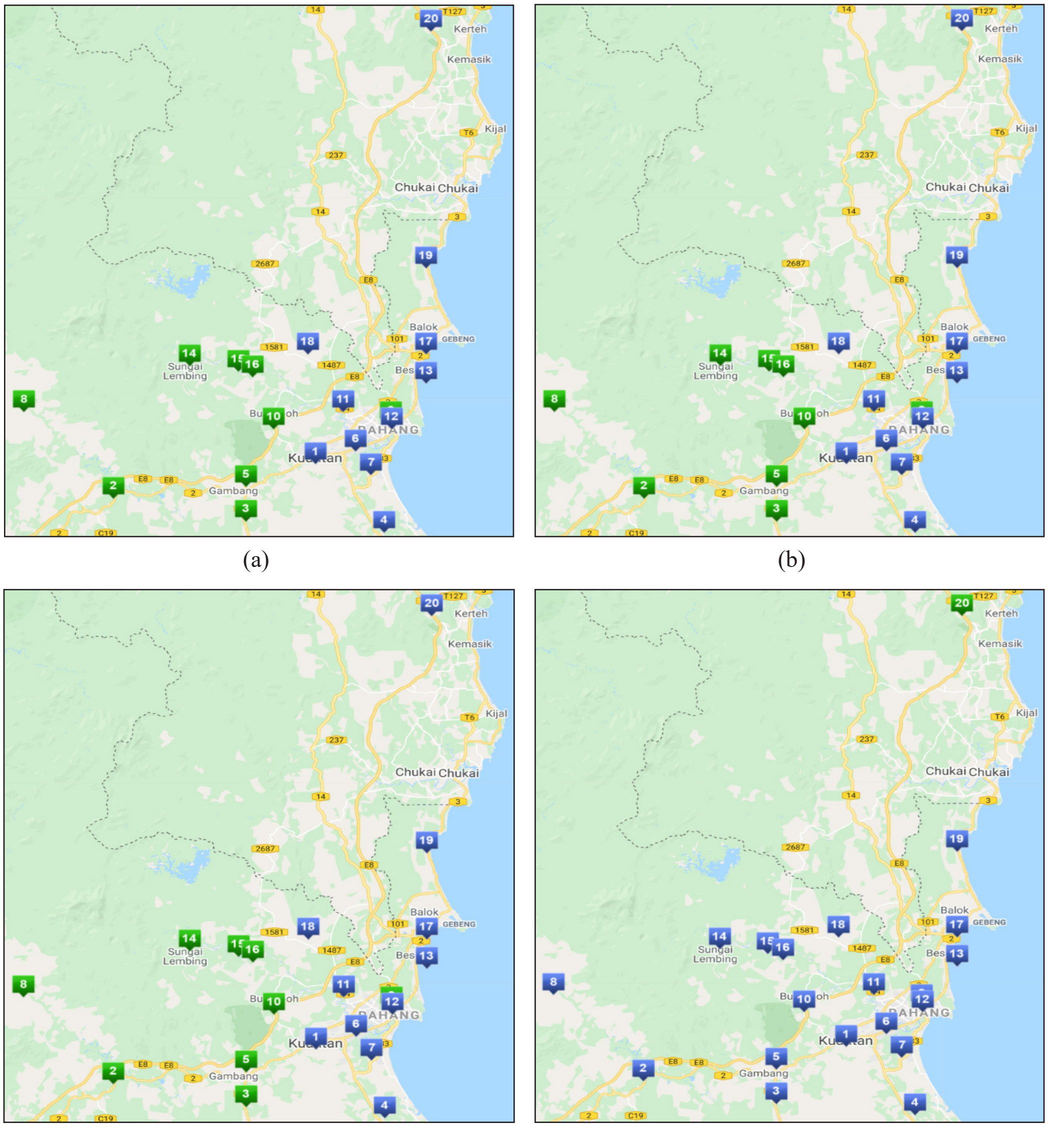
(b)

(c)

(d)

Figure 5. The identified regionalised homogeneous rainfall regions using (a) ALAH, (b) CLAH, (c) SLAH, (d) WMVAH regionalisation algorithm based on λ_2

By comparing the regionalisation accuracy rate in Table 2, the automated agglomerative hierarchical regionalisation algorithms using λ_1 yielded more adequate regionalisation results than λ_2 on average. In particular, the CLAH regionalisation algorithm uses λ_1 that yielded high regionalisation with better-fitted regionalisation analysis results and the highest regionalisation accuracy rate compared to other agglomerative hierarchical regionalisation algorithms. However, there is a need to physically relocate the misplaced rain gauge stations from the inland homogeneous rainfall regions to the coastal regions (Sahrin et al., 2018).



(a)

(b)

(c)

(d)

Figure 6. The identified homogeneous rainfall regions using (a) FKNH, (b) HKNH, (c) LKNH, (d) MKNH based on the Euclidean dissimilarity measures

The rain gauge stations that need to be relocated include the ninth, 17th, 18th, 19th, and 20th. The relocated misplaced rain gauge stations provide the statistically homogeneous after testing using the BKAD statistical test ($\mathbf{C}_1$ (inland region): $\eta_{\text{BKAD}} = 4.7681$, p-value = 0.5130; $\mathbf{C}_2$ (coastal region): $\eta_{\text{BKAD}} = 23.6609$, p-value = 1.0000). The p-value of the BKAD statistical test for $\mathbf{C}_2$ is led to 1.00. This figure (aforementioned p-value) has been rounding up from $0.\overline{9}$, which vinculum ($\overline{\ }$) represents the recurring decimals. Moreover, Table 2

Table 2
The average regionalisation accuracy rate and P-value before physically relocated the misplaced rain gauge stations

Algorithm	Dissimilarity Distance	Central Value	Regionalisation Accuracy (%)	η_{BKAD} (*P*-value)
ALAH	λ_1	Mean	28.33	$\mathbf{C}_1$: null $\mathbf{C}_2$: 2.7173 (0.8729) $\mathbf{C}_3$: null $\mathbf{C}_4$: 0.7985 (0.5913) $\mathbf{C}_5$: 23.1314 (1.0000)
		Median	33.33	$\mathbf{C}_1$: null $\mathbf{C}_2$: 2.7173 (0.8729) $\mathbf{C}_3$: 2.2541 (0.2129) $\mathbf{C}_4$: 20.3358 (1.0000)
	λ_2	Mean	29.17	$\mathbf{C}_1$: null $\mathbf{C}_2$: 0.7985 (0.5913) $\mathbf{C}_3$: null $\mathbf{C}_4$: 11.2250 (0.8328)
		Median	28.33	$\mathbf{C}_1$: null $\mathbf{C}_2$: 0.7985 (0.5913) $\mathbf{C}_3$: null $\mathbf{C}_4$: 11.2250 (0.8328)
CLAH	λ_1	Mean	79.17	$\mathbf{C}_1$: 16.5523 (1.0000) $\mathbf{C}_2$: 10.2941 (0.9508)
		Median	79.17	$\mathbf{C}_1$: 16.5523 (1.0000) $\mathbf{C}_2$: 10.2941 (0.9508)
	λ_2	Mean	21.67	$\mathbf{C}_1$: 2.7173 (0.8729) $\mathbf{C}_2$: 16.5523 (1.0000) $\mathbf{C}_3$: null $\mathbf{C}_4$: 0.7985 (0.5913) $\mathbf{C}_5$: 3.9347 (0.8414)
		Median	21.67	$\mathbf{C}_1$: 2.7173 (0.8729) $\mathbf{C}_2$: 16.5523 (1.0000) $\mathbf{C}_3$: null $\mathbf{C}_4$: 0.7985 (0.5913) $\mathbf{C}_5$: 3.9347 (0.8414)
SLAH	λ_1	Mean	13.19	$\mathbf{C}_1$: null $\mathbf{C}_2$: null $\mathbf{C}_3$: null $\mathbf{C}_4$: null $\mathbf{C}_5$: null $\mathbf{C}_6$: 8.1221 (0.8786) $\mathbf{C}_7$: null $\mathbf{C}_8$: 1.8152 (0.8442)
		Median	34.72	$\mathbf{C}_1$: null $\mathbf{C}_2$: null $\mathbf{C}_3$: 12.9962 (0.9314)
	λ_2	Mean	23.96	$\mathbf{C}_1$: null $\mathbf{C}_2$: null $\mathbf{C}_3$: null $\mathbf{C}_4$: 12.9962 (0.9314)

Table 2 (*continue*)

Algorithm	Dissimilarity Distance	Central Value	Regionalisation Accuracy (%)	η_{BKAD} (*P*-value)
		Median	23.96	$\mathbf{C}_{1}$: null $\mathbf{C}_{2}$: null $\mathbf{C}_{3}$: null $\mathbf{C}_{4}$: 32.6502 (1.0000)
WMVAH	λ_1	Mean	40.28	$\mathbf{C}_{1}$: 3.0098 (0.7224) $\mathbf{C}_{2}$: 16.5523 (1.0000) $\mathbf{C}_{3}$: 6.6316 (0.9640)
		Median	40.28	$\mathbf{C}_{1}$: 3.0149 (0.7245) $\mathbf{C}_{2}$: 16.5523 (1.0000) $\mathbf{C}_{3}$: 2.1053 (0.6921) $\mathbf{C}_{4}$: 0.7985 (0.5913) $\mathbf{C}_{5}$: null $\mathbf{C}_{6}$: null
	λ_2	Mean	27.08	$\mathbf{C}_{1}$: 16.5523 (1.0000) $\mathbf{C}_{2}$: 6.6316 (0.9640) $\mathbf{C}_{3}$: null $\mathbf{C}_{4}$: 2.7173 (0.8729)
		Median	21.67	$\mathbf{C}_{1}$: 16.5523 (1.0000) $\mathbf{C}_{2}$: 0.7985 (0.5913) $\mathbf{C}_{3}$: 3.9347 (0.8414) $\mathbf{C}_{4}$: null $\mathbf{C}_{5}$: 2.7173 (0.8729)
FKNH	Euclidean Distance	Mean	95.83	$\mathbf{C}_{1}$: 22.3519 (1.0000) $\mathbf{C}_{2}$: 5.0093 (0.3272)
		Median	95.83	$\mathbf{C}_{1}$: 22.3519 (1.0000) $\mathbf{C}_{2}$: 5.0093 (0.3272)
HKNH	Euclidean Distance	Mean	95.83	$\mathbf{C}_{1}$: 22.3519 (1.0000) $\mathbf{C}_{2}$: 5.0093 (0.3272)
		Median	95.83	$\mathbf{C}_{1}$: 22.3519 (1.0000) $\mathbf{C}_{2}$: 5.0093 (0.3272)
LKNH	Euclidean Distance	Mean	95.83	$\mathbf{C}_{1}$: 22.3519 (1.0000) $\mathbf{C}_{2}$: 5.0093 (0.3272)
		Median	95.83	$\mathbf{C}_{1}$: 22.3519 (1.0000) $\mathbf{C}_{2}$: 5.0093 (0.3272)
MKNH	Euclidean Distance	Mean	95.83	$\mathbf{C}_{1}$: 22.3519 (1.0000) $\mathbf{C}_{2}$: 5.0093 (0.3272)
		Median	54.17	$\mathbf{C}_{1}$: 16.5523 (1.0000) $\mathbf{C}_{2}$: 10.2941 (0.9508)

**Note*: null represents the hypothesis testing is invalid. This is because the $\mathbf{C}_{k}$ merely comprises a single rain gauge station.

reveals that arithmetic means and median as the central value of the regionalised features do not affect the performance of regionalisation accuracy rates, except for ALAH that uses λ_1 and λ_2, SLAH uses λ_1, WMVAH uses λ_2 and MNKH regionalisation algorithms. In particular, the use of arithmetic means and median as the central value of the regionalised features for CLAH that uses λ_1 and λ_2, SLAH uses λ_2, WMVAH uses λ_1, FKNH, HKNH and LKNH always yield similar regionalisation accuracy rates.

On the other hand, the regionalisation analysis results that use FKNH [Figure 6(a)], HKNH [Figure 6(b)], and LKNH [Figure 6(c)] regionalisation algorithms are superior to MKNH and automated agglomerative hierarchical regionalisation algorithms. FKNH, HKNH, and LKNH regionalisation algorithms yielded higher regionalisation accuracy rates than MKNH regionalised algorithms, as presented in Table 2. In addition, the regionalisation analysis results use FKNH, HKNH and LKNH regionalisation algorithms to relocate one misplaced rain gauge station. In particular, the ninth rain gauge station was misplaced in the statistically homogeneous inland region; therefore, the ninth rain gauge station was relocated in the statistically homogeneous coastal region. Furthermore, the BKAD statistical test revealed the statistically significant homogeneity for coastal ($\mathbf{C}_1$: $\eta_{\text{BKAD}} = 4.7681$, p-value=0.5130) and inland regions ($\mathbf{C}_2$: $\eta_{\text{BKAD}} = 23.6609$, p-value=1.0000) after relocating the misplaced rain gauge station. Hence, this study concluded that FKNH, HKNH, and MKNH regionalisation algorithms are superior to other agglomerative hierarchical and non-hierarchical regionalisation algorithms with prior knowledge regarding the predetermined *K* number of homogeneous clusters. This study recommended using different methods, such as elbow, average silhouette, and gap statistics, in determining the optimal number of homogeneous regions when there is a lack of prior knowledge regarding the predetermined *K* number of homogeneous clusters.

CONCLUSION AND FUTURE WORK

This study evaluated the effectiveness of automated agglomerative hierarchical and non-hierarchical regionalisation algorithms to identify the significant homogeneous rainfall regions for the rain gauge stations located in the Kuantan district. The automated agglomerative hierarchical regionalisation algorithms employed are based on integrating several agglomerative hierarchical regionalisation algorithms, including ALAH, CLAH, SLAH, WMVAH, and the AU statistical test. Meanwhile, the non-hierarchical regionalisation algorithms include FKNH, HKNH, LKNH, and MKNH regionalisation algorithms. In addition, the dissimilarity measures applied on the automated agglomerative hierarchical regionalisation algorithms are the centred and uncentred inverse correlation coefficients. In contrast, the dissimilarity measures applied on the non-hierarchical regionalisation algorithms are the Euclidean distance. This study used the 20 historical monthly rainfall time series data recorded from the rain gauge stations located in the Kuantan district that comprises the coastal and inland regions with distinct spatial and temporal variability to consolidate the effectiveness of the automated agglomerative hierarchical and non-hierarchical regionalisation algorithms. The analysis results show that FKNH, HKNH and LKNH are the superior non-hierarchical regionalisation algorithms compared to MKNH regionalisation algorithms. Furthermore, FKNH, HKNH, and LKNH regionalisation algorithms yielded a similar regionalisation accuracy rate.

In summary, this study proposes another regionalised feature set, which can be formed from the combination of geographical insights and the characteristics of rainfall distribution. This study obtained high regionalisation accuracy rates after the features set was applied in the non-hierarchical regionalisation algorithms. Based on the regionalisation results, the reliability and accuracy in assessing the risk of extreme hydro-meteorological events for the Kuantan district can be improved. Furthermore, the regional quantile estimates can provide a more accurate estimation compared to at-site quantile estimates. Finally, this study suggests extending this work in the future by employing other moments bases, such as L-moments, LQ-moments, PL-moments, and TL-moments, in describing the probability distribution of the historical monthly rainfall time-series data that are used as regionalisation features.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the Department of Irrigation and Drainage Malaysia to provide the historical monthly rainfall time series data employed in this study. The authors would also like to acknowledge the scribble maps developers for providing the software to create maps in this study. A word of appreciation also goes to the reviewers for their constructive comments and suggestions on this paper. Moreover, the authors would like to extend their sincere appreciation to Universiti Malaysia Pahang and the Ministry of Education Malaysia for providing the research funds of this study, namely the UMP internal grants (RDU1703184) and the Fundamental Research Grant Scheme, FRGS/1/2018/ STG06/UMP/02/16 (RDU190134).

REFERENCES

Ahmad, N. H., Othman, I. R., & Deni, S. M. (2013). Hierarchical cluster approach for regionalisation of Peninsular Malaysia based on the precipitation amount. *Journal of Physics: Conference Series*, *423*, 1-10. https://doi.org/10.1088/1742-6596/423/1/012018

Awan, J. A., Bae, D. H., & Kim, K. J. (2014). Identification and trend analysis of homogeneous rainfall zones over the East Asia monsoon region. *International Journal of Climatology*, *35*(7), 1422-1433. https://doi.org/10.1002/joc.4066

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

Chuan, Z. L., Deni, S. M., Fam, S. F., & Ismail, N. (2020). The effectiveness of a probabilistic principal component analysis model and expectation maximisation algorithm in treating missing daily rainfall data. *Asia-Pacific Journal of Atmospheric Sciences, 56*, 119-129. https://doi.org/10.1007/s13143-019-00135-8

Chuan, Z. L., Ismail, N., Shinyie, W. L., Ken, T. L., Fam, S. F., Senawi, A., & Yusoff, W. N. S. W. (2018a). 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*, 1-10. https://doi.org/10.1088/1757-899X/342/1/012070

Chuan, Z. L., Ismail, N., Yusoff, W. N. S. W., Fam, S. F., & Romlay, M. A. M. (2018b). Identifying homogeneous rainfall catchments for non-stationary time series using TOPSIS algorithm and bootstrap k-sample Anderson darling test. *International Journal of Engineering & Technology*, *7*(4), 3228-3237.

Chuan, Z. L., Senawi, A., Yusoff, W. N. S. W., Ismail, N., Ken, T. L., & Chuan, M. W. (2018c). Identifying the ideal number *Q*-components of the Bayesian principal component analysis model for missing daily precipitation data treatment. *International Journal of Engineering & Technology*, *7*(4.30), 5-10. https://doi.org/10.14419/ijet.v7i4.30.21992

Dash, M., & Liu, H. (2003). Feature selection for clustering. In T. Terano, H. Liu & A. L. P. Chen (Eds.), *Knowledge discovery and data mining current issues and new applications* (pp. 110-121). Springer. https://doi.org/10.1007/3-540-45571-X_13

Forgy, E. (1965). Cluster analysis of multivariate data: Efficiency versus interpretability of classification. *Biometrics*, *21*(3), 768-769.

Guttman, N. B. (1993). The use of L-moments in the determination of regional precipitation climates. *Journal of Climate*, *6*(12), 2309-2325. https://doi.org/10.1175/1520-0442(1993)006<2309:TUOLMI>2.0.CO;2

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

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *28*(1), 100-108. https://doi.org/10.2307/2346830

Lloyd, S. P. (1982). Least square quantization in PCM. *IEEE Transactions on Information Theory*, *IT-28*(2), 129-137. https://doi.org/10.1109/TIT.1982.1056489

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Cam & J. Neyman (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (pp. 281-297). University of California Press.

Ngongondo, C. S., Xu, C. Y., Tallaksen, L. M., Alemaw, B., & Chirwa, T. (2011). 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), 939-955. https://doi.org/10.1007/s00477-011-0480-x

Nnaji, C. C., Mama, C. N., & Ukpabi, O. (2014). Hierarchical analysis of rainfall variability across Nigeria. *Theoretical and Applied Climatology*, *123*(1-2), 171-184. https://doi.org/10.1007/s00704-014-1348-z

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

Sahrin, S., Ismail, N., & Alias, N. E. (2018). Regional frequency analysis of Peninsular Malaysia using L-moments. *Far East Journal of Mathematical Sciences*, *103*(8), 1379-1398. https://dx.doi.org/10.17654/MS103081379

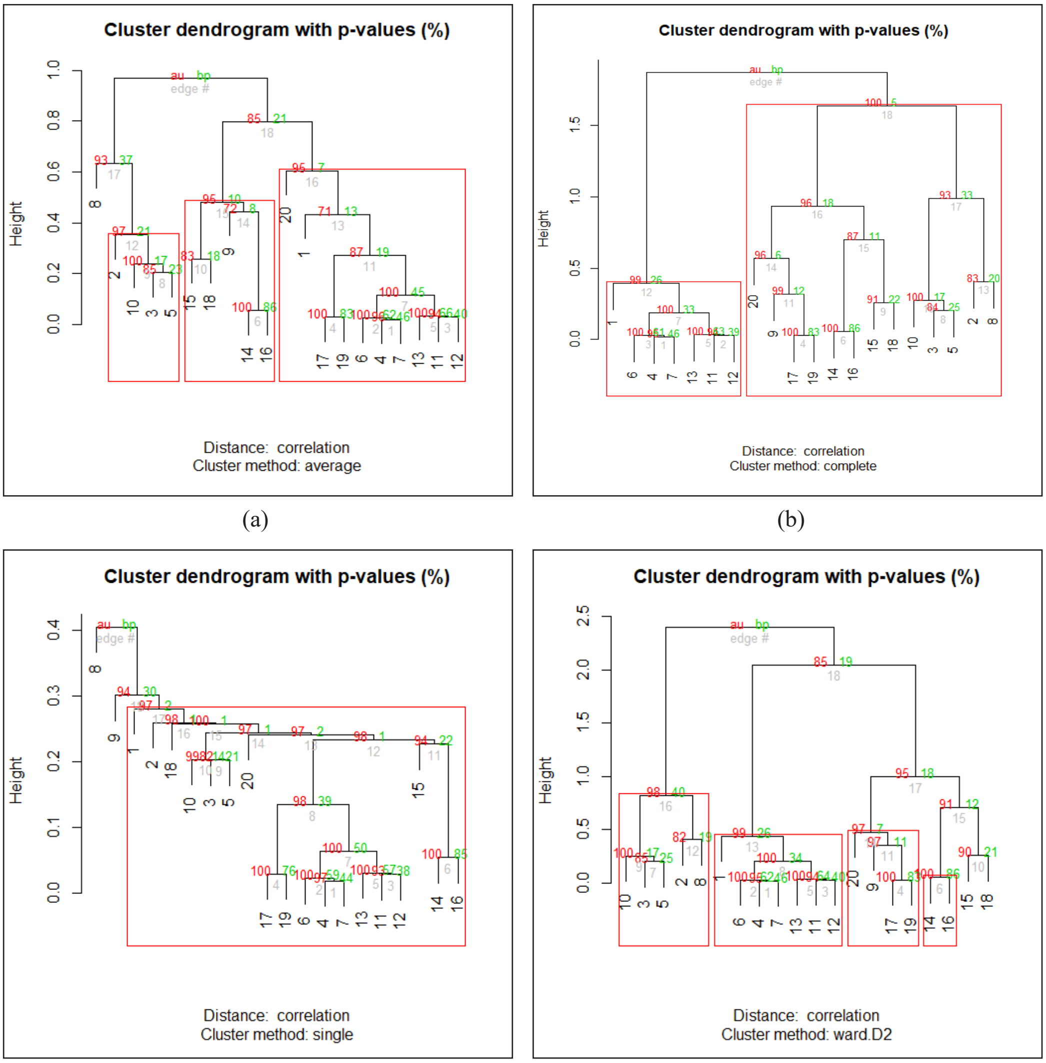
Scholz, F. W., & Stephens, M. A. (1986). K-sample Anderson-Darling tests. *Journal of the American Statistical Association*, *82*(399), 918-924. https://doi.org/10.1080/01621459.1987.10478517

Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. *Systematic Biology*, *51*(3), 492-508. https://doi.org/10.1080/10635150290069913

Tan, P. N., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*. Pearson Addison Wesley.

Terassi, P. M. D. B., & Galvani, E. (2017). Identification of homogeneous rainfall regions in the Eastern watersheds of the State of Paraná, Brazil. *Climate*, *5*(3), 1-13. https://doi.org/10.3390/cli5030053

APPENDICES



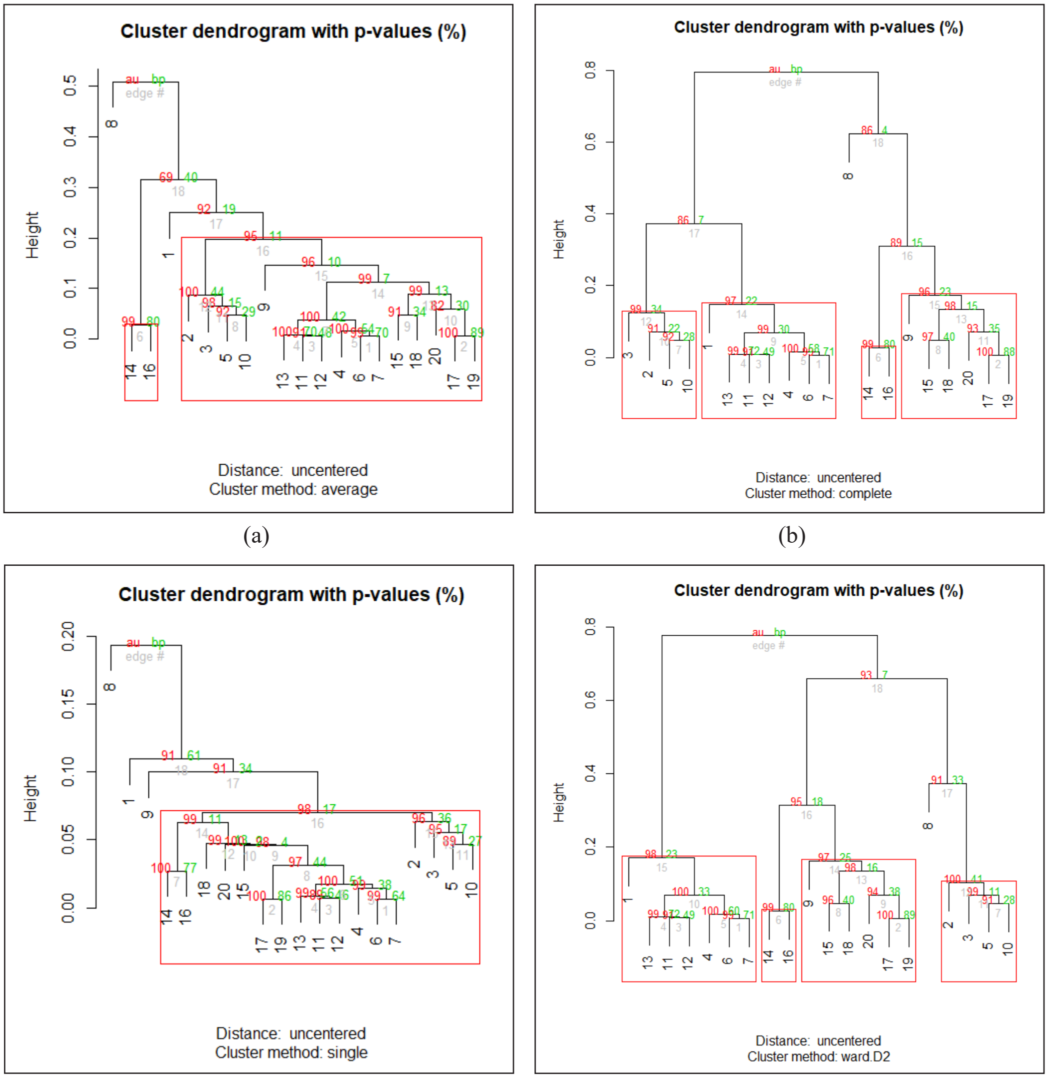
(a)

(b)

(c)

(d)

Figure A1. The dendrograms for the identified regionalised homogeneous rainfall regions using (a) ALAH, (b) CLAH, (c) SLAH, and (d) WMVAH regionalisation algorithm based on λ_1



(a) (b)

(c) (d)

Figure A2. The dendrograms for the identified regionalised homogeneous rainfall regions using (a) ALAH, (b) CLAH, (c) SLAH, and (d) WMVAH regionalisation algorithm based on λ_2