A comparison study between Doane's and Freedman-Diaconis' binning rule in characterizing potential water resources availability

A comparison study between Doane's and Freedman-Diaconis' binning rule in characterizing potential water resources availability

Zun Liang Chuan[1], Wan Nur Syahidah Wan Yusoff[1], Mohd Khairul Bazli Mohd Aziz[1], Azlyna Senawi[1] and Tan Lit Ken[2,3]

[1] Centre for Mathematical Sciences, Universiti Malaysia Pahang, Lebuhraya Tun Razak, 26300 Gambang, Kuantan, Pahang, Malaysia
[2] Takasago Thermal/Environmental Systems Laboratory, Malaysia-Japan International Institute of Technology, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia
[3] Department of Computer Science, Faculty of Knowledge Engineering, Tokyo City University, 1-28-1, Tamazutsumi, Setagaya-ku, Tokyo, 158-8557, Japan

Email: chuanzl@ump.edu.my

Abstract. One of the primary constraints for development and management of water resources is the spatial and temporal uncertainty of rainfall. This is due to the stability and reliability of water supply is dynamically associated with the spatial and temporal uncertainty of rainfall. However, this spatial and temporal uncertainty can be assessed using the intensity entropy (IE) and apportionment entropy (AE). The main objective of this study is to investigate the implications of the use of Doane's and Freedman-Diaconis' binning rule in characterizing potential water resource availability (PWRA), which the PWRA is assessed via the standardized intensity entropy (IE') against the standardized apportionment entropy (AE') scatter diagram. To pursue the objective of this study, the daily rainfall data recorded ranging from January 2008 to December 2016 at four rainfall monitoring stations located Coastal region of Kuantan District Pahang are analyzed. The analysis results illustrated that the use of Doane's binning rule is more appropriate than Freedman-Diaconis' binning rule. This is due to the resulted PWRA characteristics using Doane's binning rule is relatively consistent with practical climate such that the study region is experiencing poor-in-water zone with less amount and high uncertainty of rainfall during the Southwest Monsoon, while abundant and perennial rainfall during the Northeast Monsoon. Furthermore, the use of Doane's binning rule is more advantages compared to the Freedman-Diaconis' binning rule with the abstraction of computational cost and time.

1. Introduction

Water intimately linked to climate is a vital natural capital for ecosystem services, including freshwater supply, irrigation, hydroelectricity generation, recreation and commercial activities. Since past decades, climate change disrupts the water cycle have massively affected the amount, duration and distribution of hydro-meteorological variables such as rainfall [1]. Consequently, the uncertainty about the occurrence of rainfall emerges as one of the primary constraints for development and management of water resources. Therefore, assessing and understanding the characteristics of

potential water resource availability (PWRA) is indeed in ensuring current and future water availability to meet continuing demands.

In practical perspectives, the main interested features in assessing PWRA are the uncertainty of the occurrence of rainfall associated with time and space. This is due to the stability and reliability of water supply are dynamically associated with the spatial and temporal uncertainty of rainfall. In particular, the water supply is stable and reliable when there is low uncertainty of spatial and temporal of rainfall, respectively. Based on the literatures [2-6], coefficient of variation (CV) is the well-known measurement metrics frequently used to quantify the spatial and temporal uncertainty of rainfall across a broad geographical region of the world. However, there are very limited number of Malaysia's studies focus on the spatial and temporal uncertainty of rainfall, particularly for the East-Coast Economic Region (ECER).

On the other hand, the use of the entropy theory (ET) in quantifying the spatial and temporal uncertainty of rainfall also has been proposed [1,7,8]. Based on statistical theory and empirical evidence [9], the use of the ET is more powerful and appropriate in quantifying the uncertainty of rainfall compare to CV. In real life, the distribution of historical rainfall data is high positively skewed. However, CV is merely appropriate to quantify the uncertainty for the normal and lognormal distributed data sets and not robust to outliers, leading this measurement metric is inappropriate in quantifying the spatial and temporal uncertainty of rainfall. Conversely, the ET is more appropriate in quantifying spatial and temporal uncertainty of rainfall as this measurement metric can be fitted for any statistical distribution.

However, the challenging in assessing the characteristics of PWRA using the ET is to determine the optimum number of bins for computing the probability mass function associated the class intervals, with the abstraction of the computation cost and time. Therefore, the main objective of this study is to investigate the implications of the use of Doane's and Freedman-Diaconis' binning rule in characterizing PWRA, which the PWRA is assessed via the standardized intensity entropy (IE') against the standardized apportionment entropy (AE') scatter diagram. To pursue the objective of this study, the rest of this paper is organized as follows. Section 2 presented the description of methodology, including the selection of rainfall monitoring stations and Shannon information entropy, while Section 3 illustrated the analysis results and discussion. Finally, the concluding remarks and future work are rendered in Section 4.

2. Methodology

2.1 Selection of rainfall monitoring stations

Kuantan district in Pahang covering 1630 km^2 catchment area located at ECER is one of the unique economic development region with abundant nature and agriculture resources. One of the main tributaries in this district is Kuantan River Basin, which irrigates the major rural, agriculture, urban and industrial area [10-11]. This tropical rainforest climate's district is experiencing two different seasons annually, namely dry and hot season during the Southwest monsoon (late May to September) and rainy season during the Northeast monsoon (November to March). In specific, this district is exposed to the risk of water supply disruption due to the water scarcity of Kuantan River Basin during dry and hot season, and the occurrence of natural disasters such as flood during rainy season, resulting in a massive impact on the society and economy. Therefore, understanding and characterizing the spatial and temporal uncertainty is highly important.

In this study, daily historical rainfall data from four selected monitoring stations located Coastal Region in Kuantan River Basin [10] as illustrated in figure 1 were used to investigate the implications of the use of Doane's and Freedman-Diaconis' binning rule in characterizing PWRA. The period of rainfall time series selected is range 01 January 2008 to 31 December 2016, which this time series data sets are acquired from Department of Irrigation and Drainage Malaysia. The main features of rainfall monitoring stations and quantitative statistics of daily rainfall respect to month are provided in table 1.

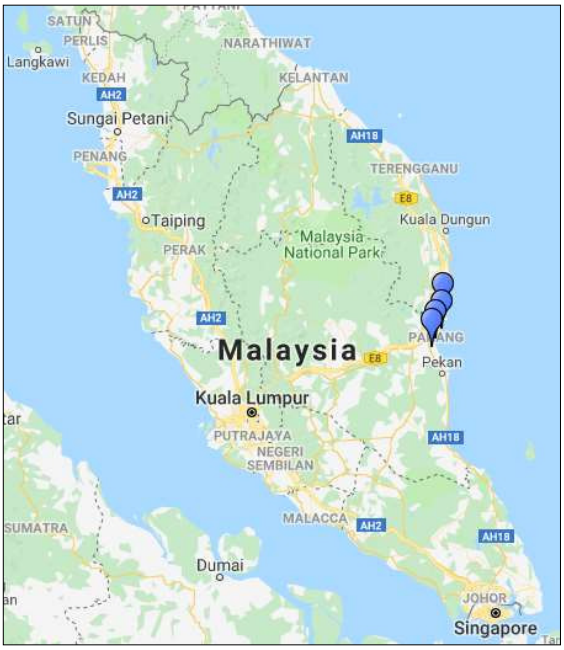

Figure 1. Location of four rainfall monitoring stations in the Kuantan River Basin.

Table 1. The main features of rainfall monitoring stations and quantitative statistics of daily rainfall respect to month for four rainfall monitoring stations in the Kuantan River Basin.

Features			Rainfall monitoring stations			
			Kampung Sungai Soi	**Pejabat JPS Negeri Pahang**	**Ladang Jeram**	**Kampung Sungai Ular**
Rainfall monitoring stations characteristics	**Abbreviation**		St1	St2	St3	St4
	Station ID		3732021	3833002	3833004	4033002
	Latitude		03°43'50"	03°48'30"	03°53'40"	04°03'00"
	Longitude		103°18'00"	103°19'45"	103°23'00"	103°23'40"
	Elevation (meters)		13	9	-32768	10
Average of daily rainfall (milimeter) ± coefficient of variation (%)	**January**	**(Jan)**	8.95 ± 63.93	7.43 ± 56.50	8.48 ± 59.66	7.18 ± 77.00
	February	**(Feb)**	2.82 ± 170.72	3.45 ± 125.79	3.64 ± 123.30	3.86 ± 131.47
	March	**(Mar)**	3.98 ± 103.46	3.77 ± 111.08	4.44 ± 125.96	4.56 ± 100.68
	April	**(Apr)**	3.46 ± 86.26	3.63 ± 88.11	4.81 ± 102.26	4.07 ± 69.18
	May	**(May)**	4.84 ± 57.48	4.82 ± 89.12	6.42 ± 81.49	5.32 ± 75.12
	June	**(Jun)**	3.43 ± 58.64	4.39 ± 73.17	4.99 ± 62.41	4.26 ± 71.63
	July	**(Jul)**	4.06 ± 65.19	3.63 ± 61.14	3.40 ± 56.93	3.37 ± 48.87
	August	**(Aug)**	5.67 ± 45.34	5.84 ± 61.27	6.28 ± 58.79	4.94 ± 49.19
	September	**(Sep)**	5.64 ± 36.57	5.20 ± 49.51	5.94 ± 67.60	5.87 ± 60.01
	October	**(Oct)**	6.66 ± 50.16	5.07 ± 71.22	6.63 ± 59.61	6.93 ± 52.60
	November	**(Nov)**	8.09 ± 35.01	7.29 ± 58.35	9.02 ± 45.57	13.58 ± 53.73
	December	**(Dec)**	24.85 ± 50.87	20.22 ± 74.91	18.90 ± 85.17	20.97 ± 63.35

2.2 Shannon information entropy

Shannon [12] introduced a procedure to approximate the expected minimum number required to encode a string of symbols based on the frequency of the symbols, such that

$$\mathrm{H}(X) = -\sum_{\kappa \in X} p_\kappa \log_2 (p_\kappa) \tag{1}$$

where H represents the entropy of a discrete random variable X with probability mass function p_κ. In this study, Shannon information entropy is used to assess the spatial and temporal uncertainty of rainfall. The main reason of this entropy version is selected as it is widely applied in hydrology studies [1,8,13].

2.2.1 Apportionment Entropy

Since the temporal of rainfall occurrence are random, the temporal distribution of rainfall, also known as apportionment of rainfall plays a substantial role in characterizing PWRA. In this study, the measurement of apportionment entropy (AE) ranges from 0 to $\log_2(\theta_j)$ is used. The high value of AE leading to the high in-time PWRA with seasonal rainfall. In the literatures [8,13], AE is regularly used to measure the temporal uncertainty of monthly rainfall over a year. However, this approach is inapplicable in this study due to the limited range of historical rainfall data. In tailor this time series data sets, the time-scale are adjusted in measuring AE.

Suppose that $\mathbf{Y}_j = \{y_{ij}\};\ i = 1,2,\ldots,\theta_j;\ j = 1,2,\ldots,12$ represents the ith daily rainfall corresponding to the jth month over a year with central value $\bar{\mathbf{Y}}_j = \frac{1}{\theta_j}\sum_{i=1}^{\theta_j} y_{ij}$, and r_{ij} represents the daily rainfall. By equating the $p_\kappa = \frac{r_{ij}}{R_j};\ R_j = \sum_{i=1}^{\theta_j} r_{ij}$ in equation (1), the AE_l corresponding to lth year with $l = 1,2,\ldots,m$ is resulted. Subsequently, the standardized AE, AE' as the abscissa of two-dimensional scatter diagram is computed using equation (2).

$$\mathrm{AE}'_j = \frac{\overline{\mathrm{AE}_l}_j - \overline{\overline{\mathrm{AE}}}_j}{S_{\overline{\mathrm{AE}}j}} \tag{2}$$

which the central value, $\overline{\overline{\mathrm{AE}}} = \frac{1}{m}\sum_{l=1}^{m}\overline{\mathrm{AE}_l}$ and standard deviation, $S_{\overline{\mathrm{AE}}} = \frac{1}{m-1}\sum_{l=1}^{m}\left(\overline{\mathrm{AE}_l} - \overline{\overline{\mathrm{AE}}}\right)^2$.

2.2.2 Intensity Entropy

In the context of hydrology, the intensity entropy (IE) is a semi-infinite measurement, $0 \le \mathrm{IE} < \infty$, use to decipher the uncertainty of rainfall intensity. This measurement is highly associated with the distribution of rainfall. For instance, the low value of IE depicted the low uncertainty of rainfall intensity, leading to a high skewed distribution of frequency of rainfall and vice versa. Therefore, determine the number of classes is essential in computing the IE.

In this study, the Doane's and Freedman-Diaconis' binning rule is respectively applied to determine the optimum number of classes, α_j and β_j with the purpose of minimizing the computational cost and time. In mathematics,

$$\alpha_j = \left\lceil \log_2\left(2\theta_j\left|\lambda_{3j}\right| + 2\theta_j S_{\lambda_{3j}}\right) - \log_2\left(S_{\lambda_{3j}}\right)\right\rceil \tag{3}$$

$$\beta_j = \left\lceil \frac{\sqrt[3]{\theta_j}\left(\max\left(\mathbf{Y}_j\right) - \min\left(\mathbf{Y}_j\right)\right)}{2IQR\left(\mathbf{Y}_j\right)} \right\rceil \tag{4}$$

where $\lambda_{3j} = \frac{\theta_j}{\left(\theta_j - 1\right)\left(\theta_j - 2\right)} \sum_{i=1}^{\theta_j}\left(\frac{y_{ij} - \bar{\mathbf{Y}}_j}{s_{\lambda_{3j}}}\right)$ is the third moment skewness of the distribution for $\mathbf{Y}_j$ with a standard deviation of $S_{\lambda_{3j}} = \sqrt{\frac{6\theta_j - 12}{\left(\theta_j + 1\right)\left(\theta_j + 3\right)}}$, and $IQR\left(\mathbf{Y}_j\right)$ is the interquartile range of $\mathbf{Y}_j$, while $\lceil\cdot\rceil$ and $|\cdot|$ are the ceiling and absolute value functions, respectively. Thereupon, the standardized IE, IE' as the ordinate of two-dimensional scatter diagram is computed using equation (5).

$$\mathrm{IE}'_j = \frac{\mathrm{IE}_{kj} - \overline{\mathrm{IE}}_j}{S_{\mathrm{IE}_j}} \tag{5}$$

The IE is resulted by equating $p_\kappa = \frac{t_{kj}}{\mathrm{T}_j}$; $\mathrm{T}_j = \sum_{i=1}^{n} t_{kj}$, in equation (1), which t_{kj}; $k = 1, 2, \ldots, n$ represents the total frequency of daily rainfall belonging to the k th class with equidistant. Meanwhile, the average of $\overline{\mathrm{IE}} = \frac{1}{n}\sum_{k=1}^{n}\mathrm{IE}_k$ and standard deviation of $S_{\mathrm{IE}} = \frac{1}{n-1}\sum_{k=1}^{n}\left(\mathrm{IE}_k - \overline{\mathrm{IE}}\right)^2$ respect to the month, regardless the years are taking into account in this study.

3. Analysis results and discussion

To investigate the implications of the use of Doane's and Freedman-Diaconis' binning rule in characterizing PWRA, the daily time series from four selected rainfall monitoring stations located in coastal regions is used. Table 2 illustrated the α_j and β_j respectively determined using the Doane's and Freedman-Diaconis' binning rules. Based on table 2, it can be observed that the optimum number of classes determined using Freedman-Diaconis' binning rule is always higher compared to Doane's binning rule. This also indicated that the use of β_j is imposed higher computational cost and time in computing the IE rather than α_j. This is because the number of bins increase, it had required more effort in computing the frequency and the relative frequency for each bin. In consequence, the computation time also increases.

In this study, a simple tetramerous clustering approach is used in characterizing PWRA based on the IE' against the AE' scatter diagram, as shown in figures 2 and 3. There are four quadrants presented in figures 2 and 3. The points of the months in a year clustered in the first quadrant interpreted that the months have abundant and perennial rainfall. Conversely, the points in the third quadrant represented the poor-in-water zone with a low amount and high variability of rainfall in the sense of PWRA. Meanwhile, the second and fourth quadrants are respectively demonstrated rainfall is relatively rich in amount, but concentrated in time, and short but perennial rainfall. In other words, the points clustered in the second and fourth quadrants cannot be apparently demonstrated the water resources availability scale due to there is merely AE or IE is negative. However, this both quadrants

provided the sense that the storage reservoirs are needed to regulate undesirability of river or stream flows [8].

In principle, the ECER is experiencing a dry and hot season during the Southwest monsoon (late May to September) and rainy season during the Northeast monsoon (November to March). In spite of that, this principle might be no longer applicable nowadays. This fact can be consolidated based on the quantitative statistics of daily rainfall as depicted in table 1, which reflected the climate change. Table 1 presented that the St1, St2, St3 and St4 are experiencing considerably higher rainfall amount during January, May, August, September, October, November and December compared to February, March, April, Jun and July, while February and March show the highest variation compared to the rest of the month in the years. Since the results showed in figure 2 is noticeably parallel with the summary statistics in table 1, therefore this study concluded that the characterized PWRA using Doan's binning rule is more appropriate compared to the Freedman-Diaconis' binning rule, which imposed low computational cost and time. The main reason of partial unparalleled results between the summary statistics and pictorial of the characterized PWRA for figure 2 is due to the presence of outliers in data sets.

Table 2. The number of classes, α_j and β_j, corresponding to its class widths determined using the Doane's and Freedman-Diaconis' binning rule, respectively.

Binning Rule	Monitoring station	Month											
		Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Doane (α_j)	**St1**	15	14	15	15	14	14	14	15	14	15	14	14
	St2	15	15	14	15	15	15	15	14	14	14	15	14
	St3	15	15	15	15	14	15	15	14	14	15	14	15
	St4	15	15	15	15	14	15	15	14	14	14	14	15
Doane (Class width)	**St1**	15.85	3.95	7.05	5.65	6.25	4.95	5.15	8.15	7.35	9.25	9.55	29.65
	St2	11.85	6.65	5.05	7.35	6.05	7.05	7.35	5.55	7.65	5.65	9.75	18.05
	St3	11.55	8.35	8.35	6.85	8.95	9.75	5.55	7.95	5.15	13.15	12.55	43.25
	St4	11.45	8.85	8.05	6.45	6.65	9.65	4.75	6.85	6.95	7.85	14.55	36.35
Freedman-Diaconis (β_j)	**St1**	297	270	528	167	217	227	237	407	68	70	46	55
	St2	222	329	348	366	129	523	550	70	134	98	181	315
	St3	124	625	248	171	155	486	274	79	52	109	63	191
	St4	171	658	241	483	156	723	237	158	80	43	203	545
Freedman-Diaconis (Class width)	**St1**	0.80	0.20	0.20	0.50	0.40	0.30	0.30	0.30	1.50	2.00	2.90	7.50
	St2	0.80	0.30	0.20	0.30	0.70	0.20	0.20	1.10	0.80	0.80	0.80	0.80
	St3	1.40	0.20	0.50	0.60	0.80	0.30	0.30	1.40	1.40	1.80	2.80	3.40
	St4	1.00	0.20	0.50	0.20	0.60	0.20	0.30	0.60	1.20	2.50	1.00	1.00

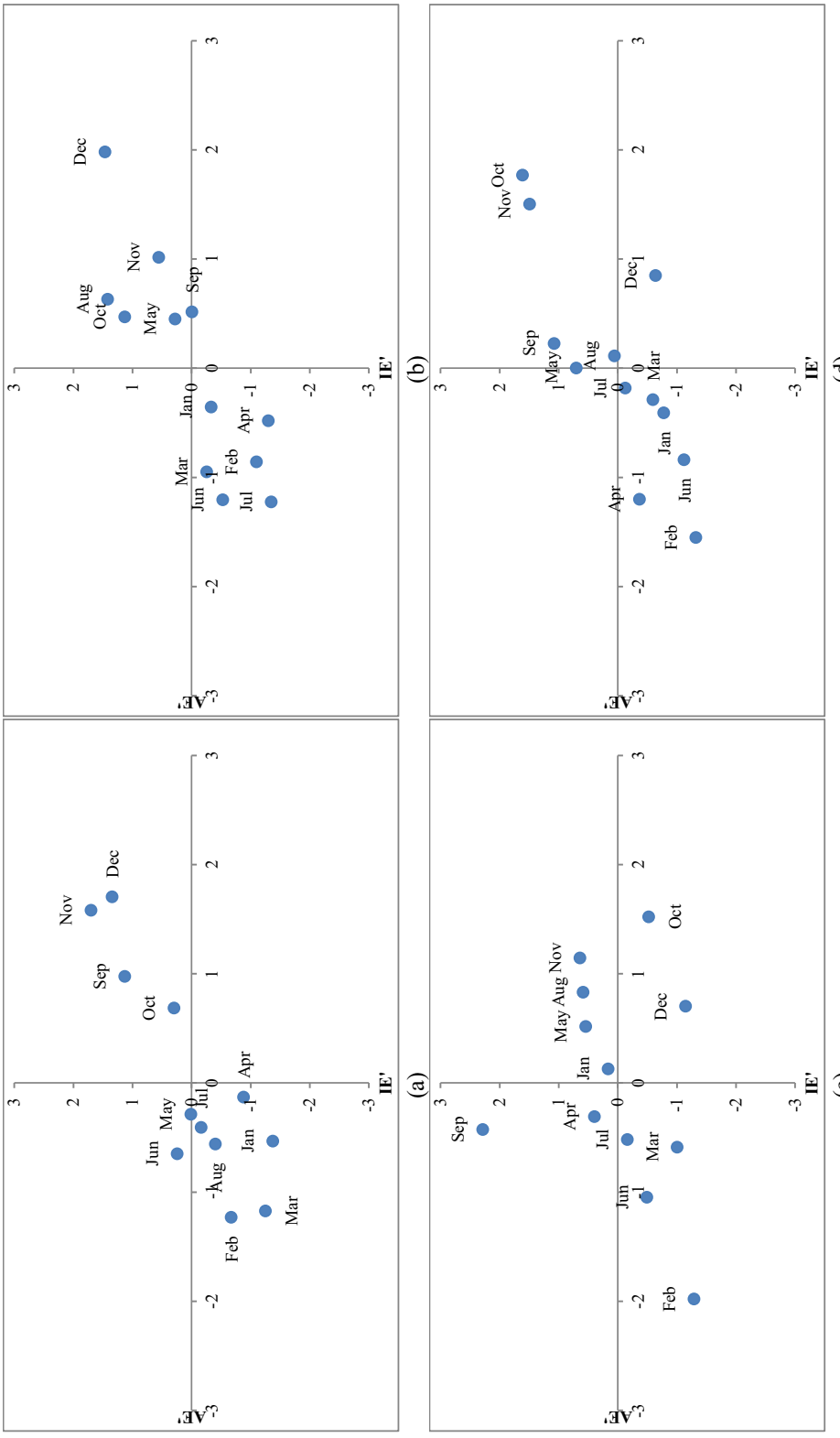

Figure 2. Characterizing PWRA using Doane's binning rule for the rainfall monitoring stations of (a) St1 (b) St2 (c) St3 and (d) St4.

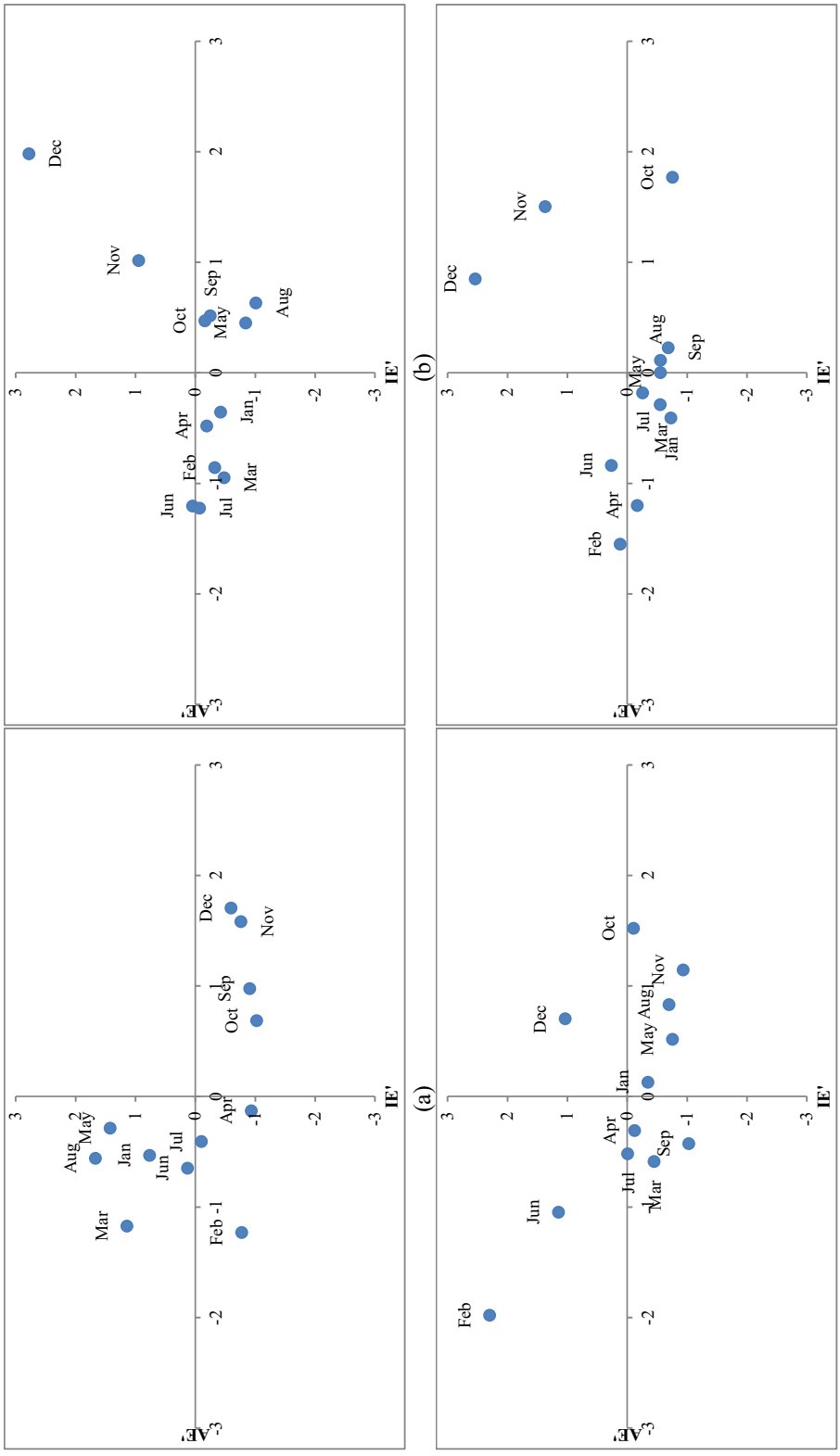

Figure 3. Characterizing PWRA using Freedman-Diaconis' binning rule for the rainfall monitoring stations of (a) St1 (b) St2 (c) St3 and (d) St4.

4. Conclusion and future work

This paper presented a study of characterizing PWRA using Doane's and Freedman-Diaconis'binning rule. The comparison of effectiveness of Doane's and Freedman-Diaconis'binning rule and its implications in characterizing PWRA also rendered in this study. Based on the analysis results, this study concluded that the Doane's binning rule is more appropriate in characterizing PWRA compared to the Freedman-Diaconis' binning rule, which imposed low computational cost and time. Although there is a partially unparalleled results between the summary statistics and tetramerous scatter diagram, however, this is an implications of the presence of outliers as classical central value and variation measurement is unrobust to outliers. In future, this study suggested to assess the spatial and temporal uncertainty focused on inland region with abstraction for development and management of water resources. This is due to the main water resources of Peninsular Malaysia are from inland rivers.

Acknowledgement

The authors gratefully acknowledge the reviewers who provided constructive comments on this paper. A word of appreciation also goes to the Department of Drainage and Irrigition (DID and Universiti Malaysia Pahang (UMP) for providing the time series data sets internal grants-RDU1703184, respectively. The authors also extend the appreciation to the Ministry of Education Malaysia for providing Fundamental Research Grant Scheme (FRGS)-RDU190134.

References

[1] Mishra A K, Özger M and Singh V P 2009 An entropy-based investigation into the variability of precipitation *Journal of Hydrology* **370**(1-4) 139-154.

[2] Fatichi S, Ivanov V Y and Caporali E 2012 Investigate interannual variability of precipitation at the global scale: is there a connection with seasonality? *Journal of Climate* **25**(16) 5512-5523.

[3] Wong C L, Liew J, Yusop Z, Ismail T, Venneker R and Uhlenbrook S 2016 Rainfall characteristics and regionalization in Peninsular Malaysia based on high resolution gridded data set *Water* **8**(11) 500. doi:10.3390/w8110500

[4] Djaman K, Sharma V, Rudnick D R, Koudahe K, Irmak S, Amouzou K A and Sogbedji J M 2017 Spatial and temporal variation in precipitation in Togo *International Journal of Hydrology* **1**(4) 97-105.

[5] Rousta I, Nasserzadeh M H, Jalali M, Haghighi E, Ólafsson H, Ashrafi S, Doostkamian M and Ghasemi 2017 Decadal spatial-temporal variations in the spatial pattern of anomalies of extreme precipitation thresholds (case study: Northwest Iran) *Atmosphere* **8**(8) 135. doi.org/10.3390/atmos8080135

[6] Sivajothi R and Karthikeyan K 2017 Spatial and temporal variation of precipitation trends in Andhra Pradesh, India *IOP Conf. Series: Materials Science and Engineering* **263** 042146. doi:10.1088/1757-899X/263/4/042146

[7] Kawachi T, Maruyama T and Singh V P 2001 Rainfall entropy for delineation of water resources zones in Japan *Journal of Hydrology* **246**(1-4) 36-34.

[8] Maruyama T, Kawachi T and Singh V P 2005 Entropy-based assessment and clustering of potential water resources availability *Journal of Hydrology* **309**(1-4) 104-113.

[9] Panichkitkosolkul W 2013 Confidence intervals for the coefficient of variation in a normal distribution with a known population mean *Journal of Probability and Statistics* **2013** 1-11.

[10] Chuan Z L, Ismail N, Shinyie W L, Ken T L, Fam S-F, Senawi A and Yusoff W N S W 2018b 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** 012070. doi:10.1088/1757-899X/342/1/012070

[11] Chuan Z L, Ismail N, Yusoff W N S W, Fam S-F, Romlay M A M 2018c 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.

[12] Shannon C E 1948 A mathematical theory of communication *The Bell System Technical Journal* **27** 379-423, 623-656.
[13] Cheng L, Niu J and Liao D 2017 Entropy-based investigation on the precipitation variability over the Hexi corridor in China *Entropy* **19**(12) 660; doi:10.3390/e19120660.