International Journal of Advanced Science and Technology
Vol. 28, No. 8s, (2019), pp. 251-259

Classification of Daily Torrential Rainfall Patterns Based on a Robust Correlation Measure

SM Shaharudin[1], N Ahmad[2]

[1]*Department of Mathematics, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, 35900 Tanjong Malim, Perak, Malaysia.*

[2]*Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia.*

Abstract

The objective of this study is to identify the main spatial distribution patterns associated with torrential rainfall days that linked to the topography of Peninsular Malaysia.This is done by applying cluster analysis on the most relevant principal directions extracted from a principal components analysis of the between day correlation. However, the characteristic of rainfall data in Peninsular Malaysia involve skewed observations which only take positive values and are skewed towards higher values. Thus, applying PCA based Pearson correlation on rainfall data set could affect cluster partitions and generate extremely unbalanced clusters. Tukey's biweight correlation is introduced to overcome the problem where the weight function down weights data values that is far from the center of the data. The findings indicate that ten rainfall patterns obtained are quite definite and clearly display the dominant role extended by the complex topography and exchange monsoons of the peninsular.

Keywords— *Tukey's biweight correlation; Principal Component Analysis; Robust correlation; Pearson correlation*

I. INTRODUCTION

Classification studies in large-scale dimensions of weather data are required to characterise synoptic and dynamic climatology patterns. Clustering techniques preceded with principal component analysis (PCA) are often combined to identify key spatial patterns in the data by reducing the number of variables for clustering cases (Shaharudin et al.,2013). A typical procedure is to first employ a PCA based T-mode correlation to reduce the dimension of the data set and k-means clustering technique to generate rainfall patterns for a particular region. This two step approach has proven to provide a clear distinction within the data, highlighting those days which presented more similarities in their spatial rainfall patterns (Fragoso et al.,2008).

A usual classification approach in identifying rainfall patterns requires the use of configuration points of entities between the rows and column of the data based on Pearson correlation matrix. Here, Pearson correlation matrix is more commonly used in the derivation of T-mode correlation to measure similarity between the daily rainfall (Romero et al.,1999, Penarrocha et al.,2002 and Sumner et al.,1995, Wickramagamage,2010).

Pearson correlation is known to perform best on normally distributed independent data and is calculated by finding the covariance of variables and dividing by the square root of the product of the variances. Here, each point are equally weighted (Chok,2008).
In analysing characteristics of rainfall pattern in Malaysia, it is important to realize that the daily rainfall variability between monsoons and regions differ (Wong, 2009) , the daily rainfall distribution is far from normally distributed (Suhaila et al.,2007) and that daily rainfall pattern is influenced by the different wet days (Chapman, 1998). Thus, applying Pearson correlation on such data might not be suitable as different days carry unequal weights and it will affect the clustering result in highlighting spatial rainfall patterns (Shaharudin et al., 2013).

In this paper, Tukey's biweight correlation is introduced as an alternative similarity measure to Pearson correlation matrix in highlighting spatial rainfall patterns in the T mode. This correlation measure are more flexible and performs well under a variety of data distributions (Hardin et al.,2007). The identification of main spatial distribution patterns is useful for hydrologist in analyzing environmental models and improves assessment on climate change.

II. DATA

In order to illustrate our new correlation matrix, we employed PCA based Tukey's biweight correlation to daily rainfall data from 75 rain gauge stations over Peninsular Malaysia. We obtained the data from Jabatan Pengairan dan Saliran (JPS) for the period 1975-2007. In this study, we focus on the occurrence of episodes on extreme rainfall event described as torrential rainfall. We have selected days that exhibited torrential characterized based on criteria described in order to retain only those days that exhibited torrential character. It was therefore necessary to choose some criteria that would lead to the establishment of a threshold, in order to allow for a clear distinction between what constitutes a day of torrential rainfall in the Peninsular Malaysia region and what does not. Area with a tropical climate with 60 mm/day is the most common threshold applied for this purpose (Shaharudin et al., 2018). By filtering days with rainfall more than 60 mm for at least 2% of overall stations, we managed to obtain 250 days and 15 rainfall stations which in turn are suffice enough to represent the main torrential centers.

III. METHODOLOGY

A. Principal Component Analysis

In climate data, the approach that used to derive the typical torrential rainfall patterns consists in subjecting the T mode analysis to principal component analysis (PCA). T mode is applied in order to analyse spatial fields in different times and it useful to extracting and reproducing the circulation types, quantifying their frequency and showing the dominat weather periods in them (Compagnucci et al.,2000). The steps involved in PCA algorithm are as follows (Samsudin et al., 2018) :

Step 1 : Obtain the input matrix
Step 2 : Calculate its Pearson correlation matrix
Step 3 : Calculate the eigenvectors and eigenvalues of the correlation matrix
Step 4 : Select the most important principal components based on cumulative percentage of total variation
Step 5 : Derive the new data set

B. Tukey's Biweight Correlation

PCA based Pearson correlation matrix might not suit all types of rainfall data particularly in the Peninsular Malaysia due to rainfall data are inherently skewed, usually to the right as such data only take positive values and tend to skewed towards higher values. In severely skewed distributions, Pearson correlation gradually loses its advantages especially for large correlation matrix. Thus, Tukey's biweight correlation in PCA is proposed to overcome the problem. The steps involved in the proposed algorithm are as follows :

Step 1 : Obtain the input matrix

Step 2 : Standardize the observation with median and mean absolute deviation (MAD), i.e.

$$x_{ij}^{*} = \frac{x_{ij} - \bar{x}}{\text{median}\left(\left|x_{ij} - \text{median}\left(x_{ij}\right)\right|\right)}$$

such that x_{ij} refer to elements in the input matrix

Step 3 : Set the breakdown point and calculate the Tukey's biweight correlation

Step 4 : Calculate the eigenvectors and eigenvalues of the correlation matrix

Step 5 : Select the cumulative percentage of variation in robust PCA

Step 6 : Derive the new data set

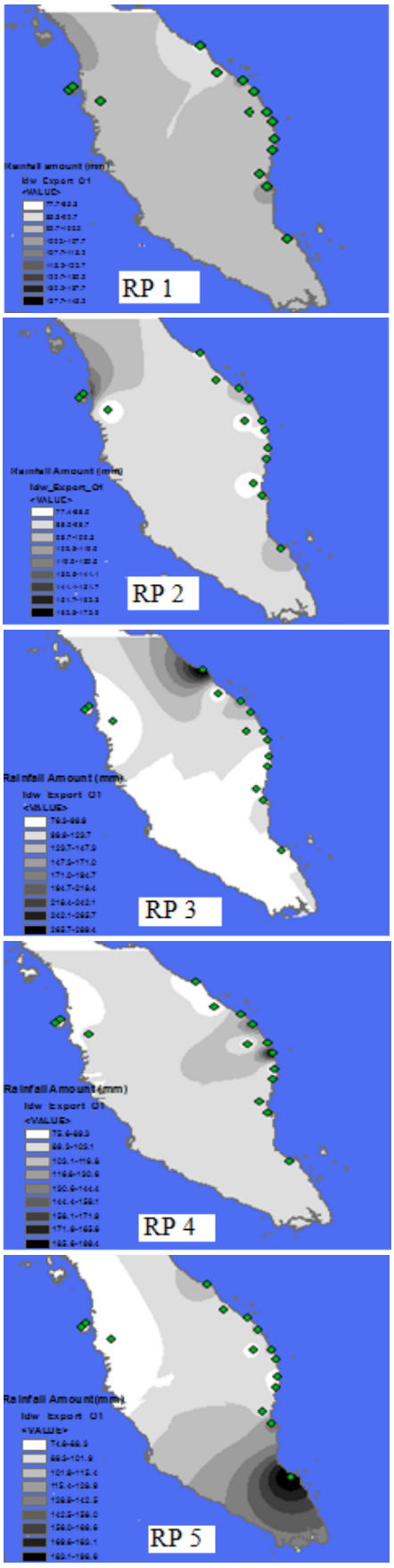
Step 7 : Apply k-means method to new data set

IV. RESULTS AND DISCUSSION

Table 1 shows the number of clusters obtained using two different approaches in PCA based correlation. PCA based Pearson correlation, produces only two clusters regardless of the cumulative percentage of variation used. Indicative of some influential observations in the data. PCA based Tukey's biweight correlation shows differentiating patterns on the number of clusters produced at different cumulative percentage of variation used. In hydrological studies particularly in identifying rainfall patterns, it is more reasonable to obtain more than two clusters to explain the various types of rainfall patterns. Thus, two cluster set is clearly are inappropriate as it masks the true structure of the data. In identifying spatial rainfall pattern, the clustering output at 70% cumulative percentage of variation on PCA-based Tukey's biweight correlation (10 clusters) is chosen respectively.

TABLE I. THE NUMBER OF CLUSTERS OBTAINED USING TWO DIFFERENT APPROACHES IN PCA BASED CORRELATION

Cum. Percentage of variance (%)	60	65	70	75	80
No. of cluster (k) using PCA	2	2	2	2	2
No. of cluster (k) using Robust PCA	12	12	10	6	2



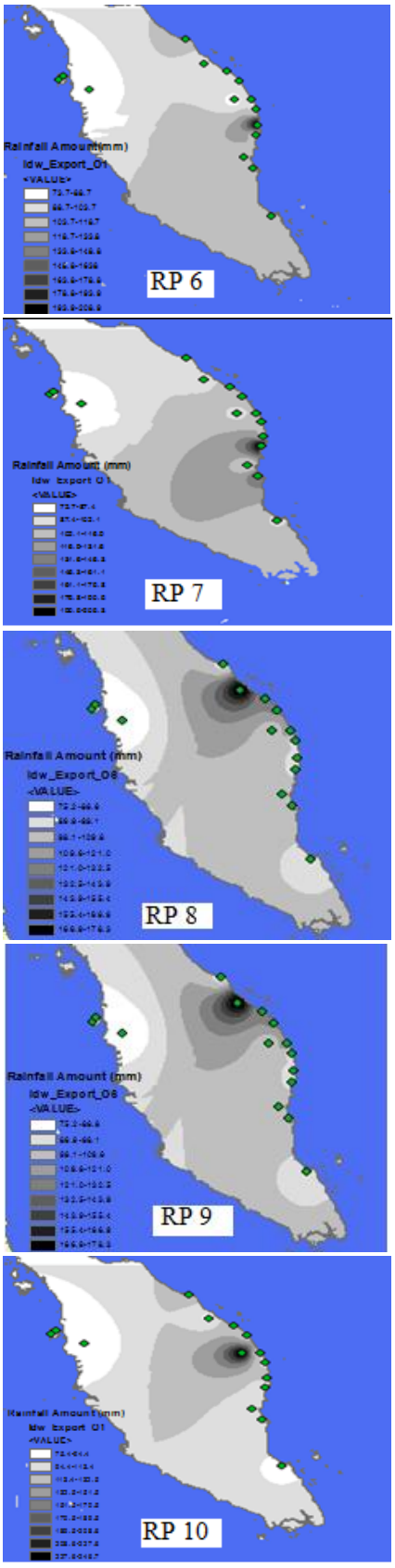


Fig. 1.Daily rainfall composites for the ten RPs obtained in the classification of torrential events

TABLE II. TABLE 1.SUMMARY OF THE TEN RAINFALL PATTERN GROUPS OBTAINED FOR DAILY TORRENTIAL RAINFALL

Rainfall Pattern	Region	Location	Days Included	Rainfall Patterns
RP 1	Northern	Pintu A. Bagan	17	Moderate torrential rainfall
RP 2		Bukit Bendera	17	
RP 3	Eastern	Kota Bahru	18	Heavy torrential rainfall
RP 4		Dungun	19	
RP 5		Kemasek	27	Heavy and intense torrential rainfall
RP 6		Kemaman	29	
RP 7		Kg. Jabi	41	
RP 8		Kg. Menerong	32	
RP 9		Endau	28	Substantial torrential rainfall
RP 10		Kuantan	22	Light torrential rainfall
			250	

*RP refer to rainfall pattern

The main features of the clustering result are discussed to verify the distinction between the clusters with respect to their significant locations and period of monsoon occurrence for the torrential rainfall patterns based on the recommended settings in the previous methodology section. In defining the spatial characteristics of torrential rainfall pattern in Peninsular Malaysia, ten clusters are obtained. These clusters are mapped out using ArcGIS software. Note that the torrential rainfall maxima locations could be clearly identified in a dark scale from the maps.

RP 1 and RP 2 exhibits moderate rainfall in the whole region in Peninsular Malaysia, with a general increase for most of the highland uplands, such as Bukit Bendera (Pulau Pinang). The maximum torrential rainfall for RP 1 occurs in the Pintu Air Bagan Air Itam (Pulau Pinang) and RP 2 in Bukit Bendera (Pulau Pinang), but the main feature of this particular pattern is the wide distribution of the torrential rainfall areas, which comprise the entire region.

RP 3 represents heavy rainfall over the east region with maximum torrential rainfall in Kota Bahru (Kelantan). In this pattern, the intensity of the torrential rainfall decreases

from east to the southwest region. Torrential rainfall occurs mostly in Kelantan due to strong influenced by the northeast monsoon and occurrence of sea breeze. Furthermore, Kota Bahru (Kelantan) is located near the coast.

RP 4, RP 5, RP 6, RP 7 and RP 8 are characterized by heavy torrential and intense rainfall occurring in the eastern region. All the patterns are located in Terengganu at different areas where RP 4 received higher rainfall in Dungun, RP 5 received heavy rainfall in Kemasek, RP 6 received maximum torrential rainfall in Kemaman, RP 7 received maximum torrential rainfall in KampungJabi and heavy torrential rainfall occurred in KampungMenerong at RP 8. Distribution of rainfall patterns for each group is significantly different due to different altitudes where the rainfalls are observed. Northwestern region received less rainfall caused by blocked by the Titiwangsa Range which possibly affects most of the rainfall stations along the western part of Peninsular Malaysia. Meanwhile, the eastern part in Peninsular Malaysia is considered as wettest area due to the strong influenced by northeast monsoon that bring heavy rainfall to east region in the period of November until March.

RP 9 represents substantial rainfalls with maximums in Endau, Mersing where the topography is defined as lowland area. Naturally, water of the rainfall will flow from high to low area. Hence, when the northeast monsoon brings heavy rainfall to that area, this make the location receives heavy rainfall. Furthermore, without ranges or mountains, the region is more likely to encounter rainfall. Due to the location concentrated close to the coast, the occurrence of sea breeze is also one of the major factors that cause this region to receive maximum rainfall. It can be seen clearly from the Figure 3, the pattern exhibits a gradual decrease from eastern to northern region.

RP 10 shows that torrential rainfall of Kuantan (Pahang) is concentrated to urban area where it is characterized by higher population density and vast human features in comparison to areas surrounding it. Normally, plants especially largest trees in urban area are difficult to find as urban areas undergo continuous built up of urban development that is within a labor market. Therefore, when northeast monsoon bring heavy rainfall in that region, it will directly receive heavy rainfall without any restriction from the largest trees.

As seen in Table 2, RP 7 pattern is significantly more frequent then the remainder RPs. This followed by KampungMenerong with RP 8. The least RPs is RP 1 and RP 2 that received less torrential rainfall and both of these patterns are located in northern region. From Table 2, it can be seen clearly that northeast monsoon was recorded highest frequency of percentage distribution of torrential rainfall occurred in each rainfall patterns and Figure 4 can illustrated that an accentuated maximum is observed in this torrential rainfall during northeast season (November to March). During this period, the winds over the east coast states of Peninsular Malaysia may reach 30 knots with strong surges of cold air from the north. This is the most substantial monsoon for all RPs. Intermonsoon season loses its relative importance in this study as torrential events are rarely observed during April and October.

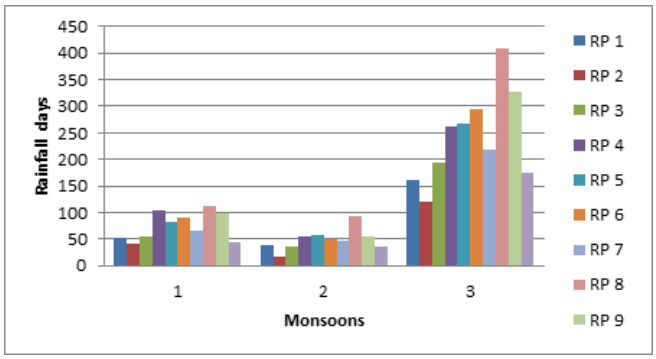


Fig. 2. Monsoons distribution for the ten rainfall patterns of torrential daily rainfall

V. CONCLUSION

PCA based Tukey's biweight correlation is a useful approach to obtain spatial patterns of torrential rainfall. Ten of RPs were obtained from k means clustering result. RP 8 is clearly associated with a higher amount of rainfall, comprising rainfall days with a torrential character. Most of the RPs pattern particularly affect the eastern region, exhibiting a strong rainfall maximum mostly in Terengganu. The majority of the location that received heavy rainfall are located in the coast and this is the one of the factor that heavy rainfall occurred in the region. All the torrential days included in these patterns occurred mostly in the northeast monsoon season. A visual inspection of those RPs confirms that the extent of region and the exposure systems induced by the complex topography are sufficiently important as to produce a clear regionalization of the rainfalls.

Acknowledgment

The authors would like to thank Universiti Pendidikan Sultan Idris for their financial funding through GPU grant Vote No. 2018-0154-101-01.

References

[1] Behera, H.S., Ghosh, A., and Mishra, S.K. (2012). A New Improved Hybridized K Means Clustering Algorithm with Improved PCA Optimized with PSO for High Dimensional Data Set. *International Journal of Soft Computing and Engineering (IJSCE)*, 2,2231-2307.

[2] Bunkers, M.J., Miller, J.R., and DeGaetano A.T.(1996), Definition of climate regions in the Northern Plains using an objective cluster modification technique, *Journal of Climate*, 9,130-146.

[3] Chapman, T.G.(1998). Stochastic modelling of daily rainfall: The impact of adjoining wet days on the distribution of rainfall amounts. *Environmental Modelling and Software*, 13,317-324.

[4] Choulakian, V. (2001). Robust q-mode principal component analysis in L_1. *Computational Statistics & Data Analysis*, 37,135-150.

[5] Compagnucci, R.H., Araneo,D., and Canziani, P.O. (2001). Principal sequence pattern analysis : A new approach to classifying the evolution of atmospheric systems. *International Journal of Climatology*, 21,197-217.

[6] Fragoso, M., and Gomes, P.T. (2008). Classification of Daily Abundant Rainfall Patterns and Associated Large-Scale Atmospheric Circulation Types in Southern Portugal. *International Journal of Climatology*, 28,537-544.

[7] Halkidi, M., Batistakis, Y., and Vazirgiannis, M.(2001). On clustering validation techniques. Journal of Intelligent Systems, 17,107-145.

[8] Hardin, J., Mitani. A., Hicks. L., and Vankoten. B.(2007). A robust measure of correlation between two genes on a microarray. BMC Bioinformatics, 8,220.

[9] Jamaludin, S., and Jemain A.A. (2007). Fitting the statistical distributions to the daily rainfall amount in Peninsular Malaysia. *Jurnal Teknologi*, 46:33-48

[10] Jamaludin, S., MohdDeni, S., Wanzin, W.Z., and Jemain, A.A. (2010). Trends in PeninsulatMalaysia Rainfall Data during the Southwest Monsoon and Northeast Monsoon Seasons. *Sains Malaysian*, 39(4),533-542.
[11] Jolliffe, I.T.(2002). *Principal Component Analysis*. (2nd ed.). New York,Inc. : Springer-Verlag.
[12] Kafadar, K.(1983). The efficiency of the biweight as a robust estimator of location. *Journal of Research of the National Bureau of Standard*, 88,105-116.
[13] Mimmack, G.M., Mason S.J., and Galpin, J.S.(2000). Choice of distance matrices in cluster analysis:Defining regions. *Journal of Climate*, 14,2790-2797
[14] Penarrocha, D., Estrela, M.J., and Millan, M.(2002). Classification of Daily Rainfall Patterns in a Mediterranean Area with Extreme Intensity Levels: The Valencia Region. *Internation Journal of Climatology*, 22,677-695
[15] Romero, R., Ramis, C., and Guijarro, J.A. (1999). Daily Rainfall Patterns in the Spanish Mediterranean Area: An Objective Classification. *International Journal of Climatology*, 19,95-112
[16] Sumner, G., Guijarro. J.A., Ramis, C. (1995). The impact of surface circulations on the daily rainfall over Mallorca. *International Journal of Climatology*, 15, 673–696
[17] Wickramagamage, P. (2010). Seasonality and spatial pattern of rainfall of Sri Lanka: Exploratory factor analysis. *International Journal of Climatology*, 30,1235-1245
[18] Wong, C.L., Venneker, R., Uhlenbrook, S., Jamil, A.B.M., and Zhou, Y. (2009). Variability of rainfall in Peninsular Malaysia. *Hydrology and earth system sciences discussions*, 6, 5471-5503
[19] Shaharudin, S.M., Ahmad, N., Yusof, F. (2013). The Comparison of T-Mode and Pearson Correlation Matrices in Classification of Daily Rainfall Patterns in Peninsular Malaysia. *Matematika*. 29 (1c), 187-194.
[20] Shaharudin, S.M., Ahmad, N., Yusof, F. (2013). Improved Cluster Partition in Principal Component Analysis Guided Clustering. *International Journal of Computer Applications*. 75(11), 22-25.
[21] Shaharudin, S.M., Ahmad, N., Yusof, F., Mohamed, N.S. (2018). Identification of Rainfall Patterns on Hydrological Simulation using Robust Principal Component Analysis. *Indonesian Journal of Electrical Engineering and Computer Science*. 11(3), 1188-1193.
[22] Samsudin, M.S., Azid, A., Khalit, S.I., Shaharudin, S.M., Lananan, F., Juahir, H. (2018). Pollution Sources Identification of Water Quality Using Chemometrics: a Case Study in Klang River Basin, Malaysia. *International Journal of Engineering & Technology*. 7(4.43), 83-89.