MATEMATIKA, 2013, Volume 29, Number 1c, 187-194
© Department of Mathematical Sciences, UTM

The Comparison of T-Mode and Pearson Correlation Matrices in Classfication of Daily Rainfall Patterns in Peninsular Malaysia

[1]Shazlyn Milleana Shaharudin, [2]Norhaiza Ahmad, [3]FadhilahYusof and [4]XenQuan Yap
[1,2,3]Department of Mathematics, Faculty of Science, Universiti Teknologi Malaysia,
81310 UTM Johor Bahru, Johor, Malaysia,
[4] Institute of Geospatial Science & Technology (INSTeG), Faculty of Geoinformation and Real Estate,
81310 UTM Johor Bahru, Johor, Malaysia.
e-mail : [1]shazlyn.milleana@gmail.com, [2]norhaiza@utm.my, [3]fadhilahy@utm.my, [4]xqyap2@live.utm.my

Abstract The aim of this study is to identify daily rainfall patterns of wet days linked to the topography of Peninsular Malaysia using two different configurations of points in the data. The data used in this study were obtained from 75 rain gauge stations in Peninsular Malaysia from the year 1975-2007. We only consider data for the period in which southwest monsoon occur from June until September yielding a total of 153 days.A typical classification approach in identifying daily rainfall patterns requires the use of configuration points of entities between the rows and column of the data based on correlation matrices. In this study, we compare effect on the cluster of daily rainfall patterns on two types of correlation matrices: T-mode correlation matrix and Pearson correlation matrix. These matrices are then used as inputs for Principal Component Analysis (PCA) to reduce the dimension of the dataset before clustering the rainfall patterns of wet days. We have found that although T-mode correlation matrix is popularly used in subtropical climate studies, it is unable to show clear classification in defining daily rainfall patterns in tropical climate data. Using Calinski and Harabasz Index, only two-rainfall pattern cluster can be identified on T-mode correlation matrix. On the other hand, Pearson correlation matrix showed three different rainfall patterns and each cluster are identified to be linked to certain topographic characteristics. These three clusters indicate that the rainfall pattern during the southwest monsoon experiencing the most heavy rain in the western part of the Peninsula, particularly in characterizing the rainfall pattern of the northwestern and western region of Peninsular Malaysia. These clusters are mapped out using ARCGIS software.

Keywords T-mode Correlation Matrix; Pearson Correlation Matrix; PCA; k-means clustering; Calinski and Harabasz Index; Southwest Monsoon; Daily rainfall pattern.

2010 Mathematics Subject Classification 62H25, 62H30

1 Introduction

The southwest monsoon season occurs annually from June to the end of September each year. The prevailing wind flow is generally southwesterly and light, below 15 knots. During southwest monsoon, the west coast of Peninsular Malaysia receives heavy rainfall as compared to the east coast area. This is because southwesterly wind blowing over the mountain range of Sumatra brings heavy rainfall to the west coast area in Peninsular Malaysia. As a result, the areas in the west coasts tend to receive more rainfall and are known as the wettest region during southwest monsoon.

Patterns for significant rainfalls tend to display similar spatial characteristics which imply that rainfall patterns are similarly highly structured and strongly linked to topography [1]. Peninsular Malaysia is made up of highlands, lowlands, drainage, coastal lines and island.The topography of Peninsular Malaysia is dominated by a mountain range, called Main Range, which runs through the middle of the peninsula up to an elevation of about 2000 meters. Areas located in higher latitude tend to receive more rainfall in contrast to the lowland areas. However, lowland areas, has the potential to receive more rainfall if the location is far from the range, hills or mountains. Range or mountains typically block areas from receiving heavy rainfall. Therefore, it can be anticipated that the resulting rainfall patterns will be strongly linked to the topography of the region and also influenced by monsoon.

The study of rainfall patterns classification is generally aimed to characterize the rainfall distribution patterns associated with heavy rain events. For instance [2], characterize the nature of precipitation regimes across Nepal and helps to elucidate the key controlling factors for spatial patterns in seasonal precipitation behavior. [3] used hierarchical and divisive cluster analysis to classify rainfall spatial time distribution pattern and also to evaluate annual and seasonal temporal pattern over Iran. Other literatures with similar objective include [4] which identified the spatial distribution patterns of heavy rainfall in the Valencia region, Spain. Most of the literatures concerning these studies are related to subtropical climate. The characteristic of the rainfall pattern in these areas are typically highly varied which depends on four seasons that occurred every year in that country. Each season is distinct by different patterns of rainfall. However, while many tropical climates have dry and wet seasons, the characteristics of rainfall patterns in these areas are usually influenced by the monsoon wind. Related studies include [5], [6], [7], [8], [9], [10] and [11].

Most of the literature review for tropical climate described above used modeling in particular regression based modeling in determining rainfall patterns but there is insufficient study relating to classification of rainfall pattern in tropical climate.This might be due to the characteristics of tropical climate which the amount of rainfall is not significantly different for each monsoon which makes it difficult to classify rainfall pattern. The aim of this study is to identify daily rainfall patterns of wet days linked to the topography of Peninsular Malaysia using two different configurations of points in the data. A typical classification approach in identifying daily rainfall patterns requires the use of configuration points of entities between the rows and column of the data based on correlation matrices. In this study, we compare effect on the cluster of daily rainfall patterns on two types of correlation matrices: T-mode correlation matrix and Pearson correlation matrix. These matrices are then used as inputs for Principal Component Analysis (PCA) to reduce the dimension of the dataset before clustering the rainfall patterns of wet days.

2 Data

The daily rainfall data from 75 stations over Peninsular Malaysia were obtained from Jabatan Pengairan dan Saliran (JPS) for the period 1975-2007. We consider the data for the period of records that ranged from June until September as this is when the southwest monsoon season occurs. The rainfall data set considered for the purpose of this study is a matrix, comprising data from 75 stations and 153 days which constitute enough data to allow for the identification of the main rainfall patterns. Therefore, the exact total number of rainfall data set is 11475 days. In this present study, a wet day is defined as a day with at least 1mm of rainfall [6]. Figure 1 shows the geographical coordinates of the stations in the study.

The rain gauge stations are grouped according to four regions namely the southwest, east, west and northwest depending on geographic coordinates. A summary of the statistics of daily rainfall amount during southwest monsoon is given in Table 1. From the table 1, we can illustrate that the highest average mean and average standard deviation of rainfall amount during southwest monsoon in Peninsular Malaysia is located in northwest region. We observed that, there are differences in the coefficient of variation between regions. The stations at the northwest and west region show the largest variability of rainfall amounts, which range from 36% to 58%. The lowest coefficient variation is found in the east station with variation less than 34%. Northwest and west region again give the largest positive skewness with starting value from 0.3 mm to 1.8 mm. The results illustrate that the shape of rainfall distribution for the stations in this two regions is much skewed compared to other regions.

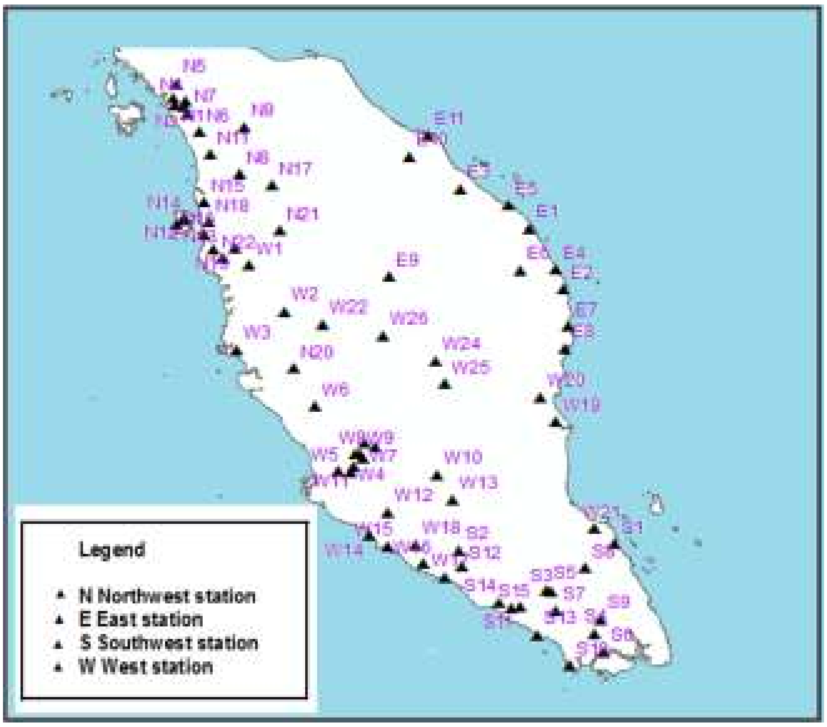

Figure 1 The location of 75 rainfall stations in Peninsular Malaysia

Table 1 Summary statistics of daily rainfall amount for each station divided by four regions

Region	Mean average (mm)	Stdev average (mm)	Range of CV (%)	Skewness (mm)
Southwest	5.5	2.3	35-49	0.3-1.5
East	5.4	2.2	34-50	0-1
West	5.5	2.5	36-58	0-1.8
Northwest	6.5	2.9	37-52	0.3-1.4

3 Methodology

In this study, we used two types of correlation matrix which is Pearson correlation matrix and T mode correlation matrix. According to [12], they use T mode correlation matrix in define daily rainfall patterns in their study to join days with similar precipitation distributions, irrespectively of the precipitation amounts. T mode correlation matrix start with spatial variance-covariance matrix and it defines as

$$C_{tt} = X'^{T}X' = \begin{pmatrix} .. & \cdots & T \\ \vdots & \ddots & \vdots \\ T & \cdots & .. \end{pmatrix} \tag{1}$$

where C_{tt}refer to spatial variance-covariance matrix and X refers to observations i.e rainfall days.
Then, the equation of T mode correlation matrix are defined as

$$C_{tt}\overrightarrow{e_t} = \lambda\overrightarrow{e_t} \tag{2}$$

where et, t = 1,2,3,…,n refer to eigenvalues of C_{tt}, λ refer to eigenvalues of C_{tt}.

Pearson's correlation coefficient between two variables is defined as the covariance of the two variables divided by the product of their standard deviations and commonly it represented by the letter r which referred to as the sample correlation coefficient or the sample Pearson correlation coefficient. We can obtain a formula for r by substituting estimates of the covariances and variances based on a sample into the formula above:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \quad (3)$$

where X_i and Y_i, i= 1,2,…,n refer to the observations in the data set i.e. rainfall days and $\bar{X}$and $\bar{Y}$are refer to the mean of each observation in the data set.

We use both of these approaches to apply in our data set and from there we make a comparison which one is better to use for further analysis.

PCA is a method to reduce the size of large data matrix to some smaller data set before proceed to further analysis. One of the important part in PCA is to determine number of component to retain and it has several rough guide to help us handle this matter. [13] recommended using 70%-75% cumulative percentage of total variation as a method to cut off the eigenvalues in large data set. The result of PCA are usually discussed in terms of component scores and loadings. In the study of identifying patterns, usually researchers were using loadings to carry out further analysis. This is the steps involved in PCA algorithm are :

Step 1 : Obtain raw data matrix
Step 2 : Calculate T mode correlation/Pearson correlation matrix
Step 3 : Employ PCA method to T mode correlation matrix /Pearson correlation matrix
Step 4 : Choosing important components using cumulative percentage of total variation
Step 5 : Take out the results of loadings in PCA to carry out further analysis

Some researcher promulgate the use of rotated components are produced better result for interpret the component loadings. To approve the statement, [14] was done some experiment with obliquely rotated component loadings which produced moderate simple structure. The precipitation patterns obtained were similar to those based on unrotated component loadings. Hence, the solution based on unrotated component loadings was finally selected to use for further analysis. The eight component loadings were extracted from the non-rotated solution in this study.

Before cluster analysis was applied, Calinski and Harabasz Index was employed to component loadings. The maximum value of the index was used to indicate the correct number of partitions in the data set. The result was obtained the best number of cluster for this analysis is three groups. Then, k means cluster analysis was performed on components loadings matrix. The aim of this procedure was to group days with similar loadings into clusters, which amounts deriving the most representative daily rainfall patterns. Euclidean distance was used in this analysis to find the index of similarity between each loading.

K means is a method to simple partition clustering technique which attempts to find a user-specified k number of clusters. The idea is to classify a given set of data into k number of disjoint clusters, where the value of k is fixed in advance. This algorithm is simple to implement and run, relatively fast, easy to adapt, and common in practice [15].

This algorithm consist of two separate phases: the first step is to define k centroids, one for each cluster. The next phase is to assign each data object to the nearest centre. Euclidean distance method is generally used to determine the distance between each data points and the cluster centres. When all the data objects are assign to each of the clusters, the cluster centres recalculation is done. The k means clustering algorithm works as follows:

Step 1 : Choose k randomly from data set as initial cluster centre
Step 2 : Calculate the distance between each data points d_i and assign each item d_i to the cluster which has the closest centroid. Recalculate the cluster centre for each cluster until convergence criteria is met.

4 Results and Discussion

Table 2 Determine number of clusters based on cumulative percentage of total variance using T mode and Pearson correlation matrix

Cumulative percentage of total variance (%)	70	75	80	85	90
No. of cluster (k) using T mode correlation matrix	2	2	2	2	2
No. of cluster (k) using Pearson correlation matrix	3	2	2	2	2

PCA analysis was then applied to both of the correlation matrix to extract the important number of components that based on the variation of the matrix set. The function of PCA method is not merely as a data reduction technique, but it will use for fundamental modes of variation of the data which are considered for the clustering process. There are many rules as a rough guide for deciding how many components to retain for the data set. Determination the number of components retained is based on the cumulative percentage of total variation is one of the famous criterions in order to separate 'signal' from 'noise' (Jolliffe, 1986) [13]. In this analysis, we take out five of the overall cumulative percentage (70%, 75%, 80%, 85%, 90%) of total variation to analyze them which fits to the data set in this study. Table 2 show the number of clusters based on cumulative percentage of total variance. We can illustrate that the different number of clusters are obtained when using two types of correlation matrix. 70% cumulative percentage of total variations came out with two numbers of clusters for T mode correlation matrix and three numbers of clusters for Pearson correlation matrix. When we increase 5% cumulative percentage of total variation, the number of clusters for T mode correlation matrix is remain the same as before while Pearson correlation matrix are decrease to two number of clusters from 75% until 90% cumulative percentage of total variation. Therefore, the best solution in total variation for this analysis is 70% based on Pearson correlation matrix where it is adequate to explain the variation of the whole data set.

We have found that by applying cluster analysis on the most relevant principal directions extracted from a PCA of the Pearson correlation matrix is more applicable to use in this study. The results were obtained three numbers of clusters and each number of clusters is represent specific characteristics based on topographic and its connection with the main rain bearing flows. Here, we discussed in detailed these patterns of significant rainfalls which are presented in map of Peninsular Malaysia.

Figure 2 shows the daily rainfall patterns of the three clusters that have been plotted. The maps of these three clusters have been drawn using the krigging method in Arcgis software. Each map has respective areas showing heavy rain during the southwest monsoon occurred. Besides that, each groups of these there clusters shown a distinct type of daily rainfall patterns that influenced by southwest monsoon. In general overview from the three clusters indicate that the southwest monsoon had the greatest impact on the northwest and west region in Peninsular Malaysia and considered as the wettest region during southwest monsoon where give the same result as [6]. Summary of the daily rainfall patterns in each group is drawn in Table 4.

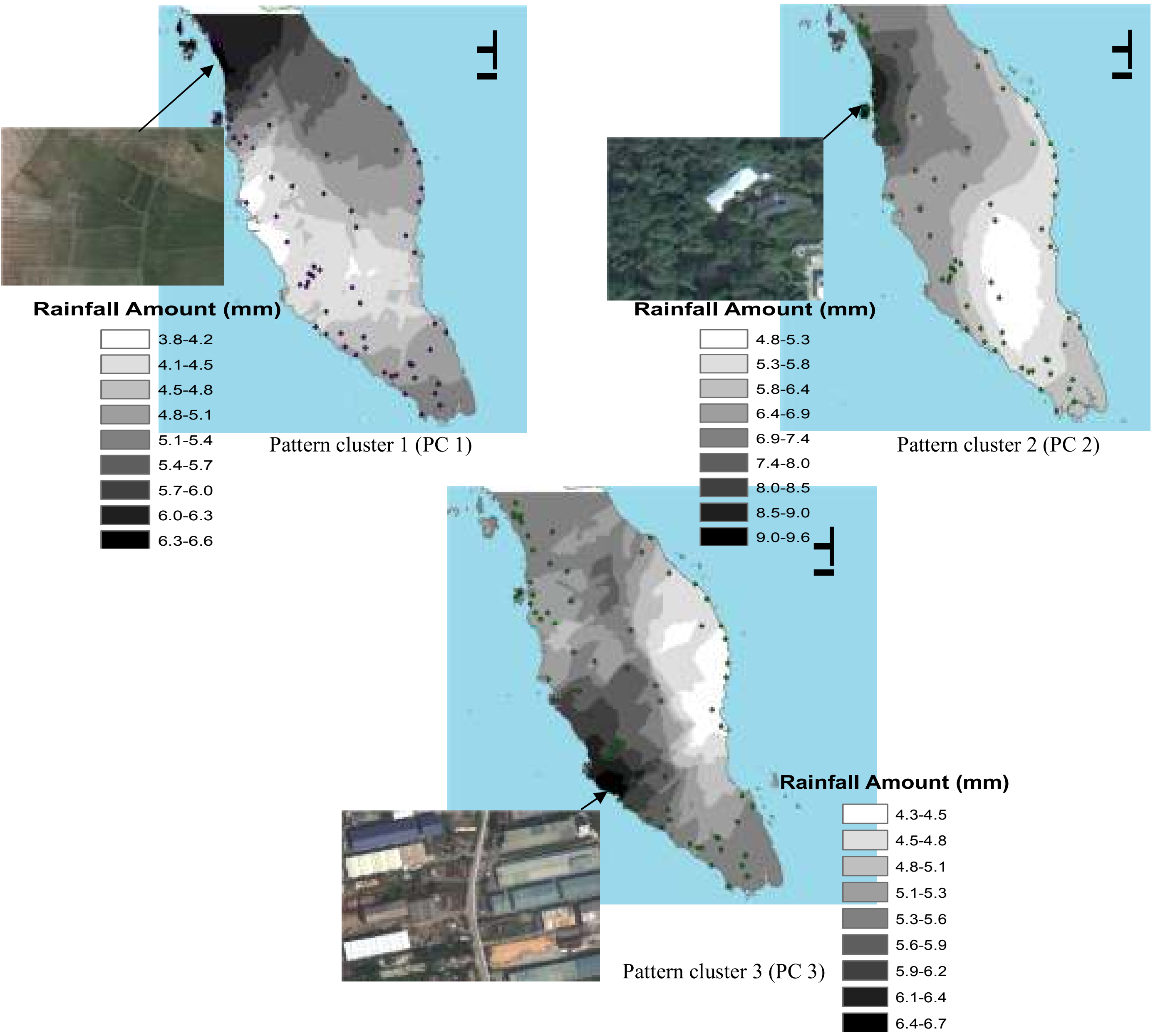

Figure 2 Daily rainfall patterns that influenced by southwest monsoon

Table4 General characteristics of daily rainfall patterns for the three clusters

Pattern Cluster	Number of days included	Daily rainfall patterns	Geographic location
1	41	Moderate rainfall in the northwest region	Lowland area–alluvial soil
2	65	Abundant and intense rainfall in Pulau Pinang	Highest location-hill area
3	47	Heavy rainfall in the west area	Lowland -urban area
Total days	153		

According to the Table 4 , cluster 1 shows the minimal number of rainfall days and the pattern of the rainfall that can be classified for PC 1 is moderate rainfall. The maximum rainfall days occurs in the northwest region especially in Padang Katong (Kangar) and Arau which located in Perlis, Kodiang and Pendang which located in Kedah. Northwest region especially in location that mention before is defined as lowland area which comprise from alluvial soil. Alluvial soil is very suitable for rice cultivation in Malaysia. Therefore, the area that had the greatest impact on moderate rainfall in northwest region is an agricultural area. Southwest monsoon brings heavy rainfall to lowland area because of nonexistence of the range or mountain in that area where the region is free from receiving heavy rainfall.

Daily rainfall pattern in PC 2 exhibits abundant and intense rainfall in Pulau Pinang. The members of cluster two is recorded the highest numbers of days that group together to become one cluster. The maximum daily rainfall is located in Klinik Bukit Bendera, Pintu Air Bagan Air Itam and KolamTakungan Air Itam which situated around the hill. Adjacent to the hill generally will receive relief rainfall also known as orographic rainfall that brought by southwest monsoon. Relief Rainfall forms where moist air is forced to rise over mountains or hills. As the air rises, it begins to cool and condenses. Clouds are formed when the water vapor condenses back to water droplet. The rainfall is started when the clouds are grown. This is the reason why cluster two becomes the region that recorded the maximum daily rainfall compared to the other clusters.

Daily rainfall pattern in PC 3 represents heavy rainfall over the west part of the region. The most area that received heavy rainfall is located in Sikamat and Port Dickson which situated in Negeri Sembilan, Petaling Jaya and Subang which situated in Selangor. It can be seen clearly from the map above in PC 3, the rainfall pattern exhibits a gradual decrease from west to east region. Topography PC 3 is similar to topography in cluster 1 which located in lowland area. The different between these clusters is PC 3 is located in 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. It is because urban areas is a continuously built up land mass of urban development that is within a labor market (metropolitan area or metropolitan region). The use of plants especially largest trees is to absorb the rainfall using their roots. Hence, when the southwest monsoon brings the rainfall, it will directly to the lowland area without having any restriction from the largest trees. This is the obvious reason why this lowland area received heavy rainfall when southwest monsoon occurred.

5 Conclusion

The weather in Malaysia is characteristics by two monsoon regimes which is southwest monsoon that occurred in June until September and northeast monsoon from November to March. The influenced by monsoon is the biggest factor in determine rainfall patterns in Asia especially in Malaysia. Generally, when southwest monsoon brings heavy rainfall, it had affected in northwest and west region and considered as the wettest region during southwest monsoon occurred. The primary goal of this study is to improve the knowledge of geographic incidence of heavy rainstorms in Peninsula, seeking to provide an objective description of their main spatial patterns. We have used two types of correlation matrix in this study which is Pearson correlation matrix and T mode correlation matrix. Method of PCA and k means clustering was then applied to both of this correlation matrix and the results are not significantly different between these two approaches. T mode correlation matrix should be better compare Pearson correlation matrix because of it use spatial variance covariance which it can capture more similarities in define spatial pattern. As was demonstrated by [16] and [1] , T mode correlation matrix is more suitable to use in define daily rainfall patterns due to the advantage of providing a primary distinction within the data, highlighting those days which presented more similarities in their spatial patterns. However, in this study the result of T mode correlation matrix not much help in obtain the best number of clusters for classify rainfall pattern. When we apply T mode formula in our data set, the values of T mode correlation matrix become too small compared with Pearson correlation matrix. It affects the result of matrix loading which automatically become very small and it has a problem in distinguishing the loading to many groups when k means method is performed on matrix loading. This method is not suitable to use in tropical climate data but we expect that the results of T mode correlation matrix could be developed and improved in the future by putting weights in T mode correlation matrix which may allow cluster analysis distinguish the loading more easily.

Acknowledgement

The authors would like to thank UniversitiTeknologiMalaysia, Malaysian Ministry of Higher Education for their financial funding through GUP grant Vote No. (Q.J130000.2626.04J09)

References

[1] Romero, R., Ramis, C., and Guijarro, J.A. Daily rainfall patterns in the Spanish Mediterranean area: an objective classification. *International Journal of Climatology*.1999. 19:95-112.

[2] Kansakar, S.R., Hannah D.M., Gerrard, J. and Rees, G. Spatial pattern in the precipitation regime of Nepal. *Internation Journal of Climatology*.2004. 24:1645-1659.

[3] SaeedSoltani* and Modarres, R. Classification of spatio -temporal pattern of rainfall in Iran using a hierarchical and divisive cluster analysis. *Journal of Spatial Hydrology*. 2006.(2):1-12.

[4] Penarrocha, D., Estrela, M.J., and Millan, M. Classification of daily rainfall patterns in a Mediterranean area with extreme intensity levels: The Valencia Region. *Internation Journal of Climatology*. 2002. 22:677-695.

[5] Jamaludin, S., and Jemain, A.A. Investigating the impacts of adjoining wet days on the distribution of daily rainfall amounts in Peninsular Malaysia. *Journal of Hydrology*.2009. 368:17-25 .

[6] Jamaludin, S., MohdDeni, S., Wanzin, W.Z., and Jemain, A.A. Trends in Peninsular Malaysia rainfall data during the southwest monsoon and northeast monsoon seasons. *Sains Malaysiana* .2010. 39(4):533-542.

[7] Mohd Deni, S., .Jemain, A.A. and Ibrahim, K.. The spatial distribution of wet and dry spells over Peninsular Malaysia.*Theory Application Climatology*. 2008. 94:163-173.

[8] Rasmusson, E.M. and Carpenter, T.H. The relationship between eastern equatorial pacific sea surface temperatures and rainfall over India and Sri Lanka. *American Meteorological Society*.1983. 111(3):517-528.

[9] Dale, W.L. The rainfall of Malaya: PartI.*J Trop Geog*.1959. 13: 23-27.

[10] Chia, L.S. An analysis of rainfall patterns in Selangor. *J TropGeog*.1968. 27:1-18.

[11] Yap, W.C. The persistence of wet and dry spells in Sungai Buloh, Selangor. *Meteor Mag*. 1973. 102: 240-245.

[12] Sumner, G., Guijarro, J.A. and Ramis, C. The impact of surface circulations on the daily rainfall over Mallorca. *International Journal of Climatology*.1995.15: 673–696.

[13] Jolliffe IT. *Principal Component Analysis*. 2nd Edition. New York,Inc. : Springer-Verlag.2002.

[14] Richmann, M.B. Rotation of principal components. *J. Climatol*. 1986. 6: 293–335.

[15] Behera, H.S., Ghosh, A. and Mishra, S.K. 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)*. 2012. 2:2231-2307.

[16] Sumner, G. Daily precipitation patterns over wales: towards a detailed precipitation climatology. *Transactions of the Institute of British Geographers*. 1996.21(1):157-176.