Hierarchical Cluster Approach for Regionalization of Peninsular Malaysia based on the Precipitation Amount

N H Ahmad¹ and I R Othman’and S M Deni³

¹,²,³Department of statistics
Faculty of computer and mathematical sciences
UiTM Shah Alam Malaysia

Email: norleehainieahmad@gmail.com

Abstract: The rainfall series from 59 homogeneous stations in Peninsular Malaysia are used in this study. The series are indicated to be multifarious, with unpredictable fluctuations from year to year and from region to region. Therefore it is more appropriate to observe all stations on spatial scales corresponding to clusters instead of the individual series. Single site occurrence analysis is inadequate for regional planning. Hence, it is important to determine regional rainfall frequencies. Hierarchical cluster analysis with 11 different similarity measures and 7 different linkages are used in this study to find homogeneous rainfall groups. Validity indices are evaluated to determine the best number of clusters in a rainfall data set. The study elucidates three homogeneous rainfall regions namely A, B and C within Peninsular Malaysia, which correspond to areas with different natural features and rain pattern. Based on this research, a hierarchical clustering method is suggested in order to group individual rainfall series into a limited set of spatial clusters.

1. Introduction

A rainy day may be defined as a day that rains more than the threshold (in this study which is 1 mm), while the intensity of rainfall may be defined as the measure of the amount of rain that falls over time. According to Malaysian Meteorological Department, [19], the characteristic features of the climate of Malaysia are of uniformed temperature, high humidity and copious rainfall. Situated in the equatorial doldrums area, it is extremely rare to have a full day with completely clear sky even during periods of severe drought. Instead, it is also rare to have a stretch of a few days with completely no sunshine except during the northeast monsoon seasons. The seasonal wind flow patterns paired with the local topographic features determine the rainfall distribution patterns over the country.

During the northeast monsoon season, the exposed areas i.e. the east coast of Peninsular Malaysia experience heavy rainfalls. On the other hand, inland areas or areas which are sheltered by mountain ranges are relatively free from this influence. It is best to describe the rainfall distribution of the country according to seasons. Alternatively, lower number of rainy days is observed in the northwestern sector of the Peninsular where minimum precipitation occurs in that part of the country. By using method the coefficient of variability [1] proved that the annual rainfall variability at the eastern half is larger than at the western half. Besides, the research done by [2] in his case study of Sungai Kayu Ara, Damansara Selangor shows that greater rainfall volume fell on developed area, compared to suburban/rural area.
The nonparametric Mann-Kendall (MK) statistical test used by [3] to display the results for both southwest monsoon and northeast monsoon seasons indicate that there are significantly decreasing trends in the frequency of wet days during the extreme events for most of the stations in Peninsular Malaysia. However, a smaller number of significant trends were found for extreme intensity. In addition, [4] in their study seeks to identify the spatial patterns and time-variability of rainfall in Peninsular Malaysia on monthly, yearly and monsoon temporal scales using Shepard's interpolation scheme. The spatial variation analysis shows that the East Coast region, which received substantially higher amounts of rainfall during the northeast monsoon, has lower spatial rainfall variability and a more uniformed rainfall distribution than other regions. A larger range of the monthly spatial variation is observed in the West Coast region. Thus, rainfall variation within Peninsular Malaysia is complex, with unpredictable rainfall fluctuations from year to year and from region to region by comparison.

Quite a lot of methods are regularly used for the regionalization of hydrologic phenomena such as rainfall, streamflow and other components of the water cycle as to determine if an area has hydrologically homogeneous regions [5]. Multivariate techniques are powerful tools in classifying meteorological data such as rainfall. Principal component analysis, factor analysis, and different types of cluster analysis and support vector machine analysis have been used to classify daily rainfall patterns and their relationship to atmospheric circulation [6], flood and drought years [7], and streamflow drought [8]. The application of multivariate methods to classify hydrologic events on a national scale was suggested by [9] while L-moments and cluster analysis was applied by [10] to determine regional rainfall climates within conterminous United States, as he defined 104 homogeneous precipitation regions there. Clusters are seen to offer several advantages over geographical regions and they are interpreted in terms of basin characteristics through the use of a multivariate linear discriminant analysis as stated by [11].

There have been many studies relating to rainfall pattern of Malaysia. It provides beneficial information to various sectors including hydrological and water resources management. Nevertheless, if there are too many rain gauge stations need to be analyzed, time consumption is the utmost constraint. However, limited studies have been done to group spatial and temporal patterns to define homogeneous rainfall regions within Peninsular Malaysia in order to abridge hydrologic calculations and diminish masses of observations and variables for frequency analysis.

2. Rainfall Data

The annual rainfall data from 59 homogeneous stations in [18] are obtained from Drainage and Irrigation Department and Malaysian Meteorological Department. Data are recorded from 1975 to 2010. There are 36 years of observation used in this study. Additionally, it is suggested by The World Meteorological Organization (WMO) to use 30-year rainfall periods for rainfall analysis. Table 1 shows the 59 stations in Peninsular Malaysia with their geographical coordinates while Figure 1 shows the map of Peninsular Malaysia along with the stations used in this study. Peninsular Malaysia consists of 13 states and the boundary of each state is also illustrated.
Figure 1: Location of rainfall gauge stations in Peninsular Malaysia

Table 1: The geographical coordinates of the rain gauge stations in Peninsular Malaysia

| No. | District                           | Long  | Lat  | No. | District                           | Long  | Lat  |
|-----|------------------------------------|-------|------|-----|------------------------------------|-------|------|
| 1   | PINTU A.BAGAN,AIR ITAM             | 100.2 | 5.4  | 31  | MERSING                           | 103.8 | 2.5  |
| 2   | ARAU                               | 100.3 | 6.4  | 32  | HOSPITAL PORT DICKSON              | 101.8 | 2.5  |
| 3   | S. K. KG. AUR GADING               | 101.9 | 4.4  | 33  | SETOR JPS SIKAMAT SEREMBAN         | 101.9 | 2.7  |
| 4   | LDG. BATU KAWAN                    | 100.4 | 5.3  | 34  | PETALING JAYA                      | 101.7 | 3.1  |
| 5   | GUAR NANGKA                        | 100.3 | 6.5  | 35  | SUBANG                             | 101.6 | 3.1  |
| 6   | KLINIK BIDAN JAMBU BONGKOK         | 103.4 | 4.9  | 36  | EMPANGAN GENTING KELANG            | 101.7 | 3.2  |
| 7   | JANDA BAIK                         | 101.9 | 3.3  | 37  | GOMBAK                             | 101.7 | 3.3  |
| 8   | JAM. SG. SIMPANGAN, ILN. EMPAT     | 102.1 | 2.4  | 38  | RUMAH PAM PAHANG TUA,PEKAN         | 103.4 | 3.6  |
| 9   | IBU BEKALAN KAHANG , KLUANG        | 103.6 | 2.2  | 39  | KUANTAN                            | 103.2 | 3.8  |
| 10  | SEK. KEB. KEMASEK                  | 103.5 | 4.4  | 40  | RUMAH PAM PAYA KANGSAR             | 102.4 | 3.9  |
| 11  | SEK. KEB. KG. JABI                 | 102.6 | 5.7  | 41  | SITIAYAN                           | 100.7 | 4.2  |
| 12  | LDG. KIAN HOE , KLUANG             | 103.3 | 2.0  | 42  | JPS KEMAMAN                        | 103.4 | 4.2  |
| 13  | DISPENSARI KROH                    | 101.0 | 5.7  | 43  | LDG BOH                            | 101.4 | 4.5  |
| 14  | STN. PEMEREKSAAN HUTAN, LAWIN      | 101.1 | 5.3  | 44  | IPOH                               | 101.1 | 4.6  |
| 15  | PEKAN MERLIMAU                     | 102.4 | 2.2  | 45  | SEK.MEN. SULTAN OMAR,DUNGUN        | 103.4 | 4.8  |
| 16  | RUMAH PENIAGA JPS. PARIT NIBONG   | 100.5 | 5.1  | 46  | GUA MUSANG                         | 101.9 | 4.9  |
| 17  | PENDANG                            | 100.5 | 5.9  | 47  | KG. MENERONG                       | 103.1 | 4.9  |
| 18  | JPS. WILAYAH PERSEKUTU             | 101.7 | 3.2  | 48  | PUSAT KESIHATAN BT.KURAU           | 100.8 | 4.9  |
| 19  | LDG. BENUT ,RENAM                 | 103.4 | 1.8  | 49  | SELAMA                             | 100.7 | 5.1  |
| 20  | LDG. SG. SABALING                  | 102.5 | 2.9  | 50  | STOR JPS KUALA TRENGGANU           | 103.1 | 5.3  |
| 21  | GENTING SEMPAH                     | 101.8 | 3.4  | 51  | KOLAM TAKONGAN AIR ITAM            | 100.3 | 5.4  |
| 22  | KG. MERANG ,SETIU                 | 102.9 | 5.5  | 52  | KLINIK BKT. BENDERA                | 100.3 | 5.4  |
| 23  | IBU BEKALAN SG. BERNAM            | 101.4 | 3.7  | 53  | TO UBAN                            | 102.1 | 5.9  |
| 24  | KG. SG. TUA                       | 101.7 | 3.3  | 54  | KOTA BHARU                         | 102.3 | 6.2  |
| 25  | SENAI                              | 103.7 | 1.6  | 55  | ALOR STAR                          | 100.4 | 6.2  |
| 26  | SEK.MEN.BKT BESAR DI KOTA TINGGI    | 103.7 | 1.8  | 56  | AMPANG PEDU                        | 100.8 | 6.2  |
| 27  | SEK.MEN.INGGERIS BATU PAHAT       | 102.9 | 1.9  | 57  | PADANG KATONG ,KANGAR              | 100.2 | 6.5  |
| 28  | PINTU KAWALAN SEMBRONG           | 103.1 | 1.9  | 58  | ABI KG. BAHRU                      | 100.2 | 6.5  |
| 29  | KLUANG                             | 103.3 | 2.0  | 59  | BAYAN LEPAS                        | 100.3 | 5.3  |
| 30  | MALACCA                            | 102.3 | 2.3  |
3. Methodology
There are three steps involved in determining the regionalization of Peninsular Malaysia based on the precipitation amount. The basic steps to develop clustering process in this study are presented in Figure 2.

![Figure 2: Steps of clustering process.](image)

3.1. Determination the optimal number of cluster

Cluster analysis for rainfall precipitation involves the grouping of a given data set into groups (clusters) such that the data points in a cluster are highly similar in explaining the pattern of that particular area such as geographical, physical, statistical or stochastic features. A number of approaches can be used to do regionalization. Hierarchical cluster analysis and L-moments regional frequency analysis were both used in [5] study regarding regional precipitation climates of Iran. Multivariate linear discriminant analysis and hypothetical ungauged basins was applied by [11] in his study of drainage basins in Britain.

Cluster and principal component analysis were being employed by [9] to collect data from that of Sweden and found that cluster analysis is suitable to use on a national scale for a country with heterogeneous hydrological regimes. His research for rivers in southern Manitoba [9] Canadian incorporates both a basin similarity measure, imbedded in the clustering algorithm, and a regional homogeneity measure to evaluate station partitionings obtained from the clustering algorithm. Meanwhile, [13] used hierarchical clusters of rainfall distribution patterns in a Mediterranean region.

According to [14] hierarchical clustering techniques are proceeded by either a series of successive mergers or a series of successive divisions. 

**Agglomerative hierarchical methods** that are used in this study begin with the individual objects. Thus, there are initially as many clusters as that of object. Most similar objects are first grouped, and these initial groups are merged accordig to their similarities. Eventually, as the similarity decreases, all subgroups are fused into a single cluster.

Since there is no prior knowledge on number of clusters for regionalization of Peninsular Malaysia, eight different clustering validity indices were used to obtain the optimal number of clusters. The eight different clustering validity indices were Silhouette index, Davies-Bouldin, Calinski-Harabasz Index, Dunn Index, C Index, Krzanowski-Lai Index, Hartigan Index and Weighted Inter-Intra Index.

The clustering validity indices applied hierarchical clustering with Euclidean distances and complete linkage where in this algorithm, it uses the largest distance between objects in the two clusters. The formula for Euclidean distance is:

\[ d_{ij} = \sum_{j=1}^{p} (x_{ij} - x_{ij})^2 \]
Besides, complete linkage is the best method for classification in this study. The complete linkage is written as follows:

$$d_{ij} = \max\{d(x_i, x_j) \mid i \in \{1, \ldots, n\}, j \in \{1, \ldots, n\}\}$$

where $rth$ and $sth$ are rows of the data matrix, $x$ that are denoted by $(x_{r1}, x_{r2}, \ldots, x_{rP})$ and $(x_{s1}, x_{s2}, \ldots, x_{sP})$ respectively. These two rows correspond to the observations on two objects for all $P$ variables. The quantity $d_{ij}$ is referred to as the squared Euclidean distance. The Euclidean distance of dissimilarity is then used in the cluster techniques. Cluster Validity Analysis Platform (CVAP) package in Matlab® is used to obtain the optimal number of cluster for rainfall series in Peninsular Malaysia.

One of the most important issues in cluster analysis is the assessment of clustering results to find the segregation that best fits the data. The term optimal clustering defined as the outcome of running a clustering algorithm that best fits the inherent partitions of the data set [15]. The measure of validity of the clusters should be such that it will be able to impose an ordering of the clusters in terms of its goodness. Table 2 shows the summary of validity indices used in this study.

| Method               | Description                                           |
|----------------------|-------------------------------------------------------|
| Silhoutte index      | A larger Silhoutte value indicates a better quality of a clustering result and largest silhouette indicates the optimal number of cluster. |
| Davies-Bouldin       | A low value indicates good cluster structures and minimum value determines the optimal number of cluster. |
| Calinski-Harabasz Index | It is the pseudo F-statistic, which evaluates the clustering solution by looking at how similar the objects are within each cluster and how well the objects of different clusters are separated and maximum value indicates optimal number of cluster. |
| Dunn Index           | Large values indicate the presence of compact and well-separated clusters and maximum value indicates optimal number of cluster. |
| C Index              | Minimal C-index indicates optimal number of cluster.  |
| Krzanowski-Lai Index | Maximum value indicates optimal number of cluster.    |
| Hartigan Index       | The estimated number of clusters is the smallest k. There will be no estimation if this condition is not satisfied. However, an elbow might indicate an optimal number of clusters. |
| Weighted Inter-Intra Index | Search forward (k=2, 3, 4 ...) and stop at the first down-tick of the index, which indicates optimal number of cluster. |

### 3.2. Hierarchical Clustering

After the optimal number of cluster was obtained in the first section, hierarchical analysis was performed using 77 combinations of 11 different similarity measures and 7 different types of linkages in order to group the data into clusters. Table 3 shows the list of linkages used to create agglomerative hierarchical cluster tree as well as the list of similarity measures in Table 4. The similarity measures indicate the pairwise distance between pairs of objects. Table 3 and Table 4 were taken from the manual of Matlab® software because hierarchical clustering can be done easily using the function under cluster analysis provided by Matlab®.
Table 3: List of linkages

| Method   | Description                                           |
|----------|-------------------------------------------------------|
| Average  | Unweighted average distance (UPGMA).                  |
| Centroid | Centroid distance (UPGMC). Y must contain Euclidean distances. |
| Complete | Furthest distance.                                      |
| Median   | Weighted center of mass distance (WPGMC). Y must contain Euclidean distances. |
| Single   | Shortest distance. This is the default.               |
| Ward     | Inner squared distance (minimum variance algorithm). Y must contain Euclidean distances. |
| Weighted | Weighted average distance (WPGMA).                    |

Table 4: List of similarity measures

| Metric   | Description                                                                 |
|----------|-----------------------------------------------------------------------------|
| Euclidean| Euclidean distance.                                                         |
| Seuclidean| Standardized Euclidean distance. Each coordinate in the sum of squares is inverse weighted by the sample variance of that coordinates. |
| Mahalanobis| Mahalanobis distance.                                                       |
| Cityblock| City block metric.                                                          |
| Minkowski| Minkowski metric.                                                           |
| Cosine   | One minus the cosine of the included angle between points (treated as vectors). |
| Correlation| One minus the sample correlation between points (treated as sequences of values). |
| Spearman | One minus the sample Spearman's rank correlation between observations treated as sequences of values. |
| Hamming  | Hamming distance, the percentage of coordinates that differ.                |
| Jaccard  | One minus the Jaccard coefficient, the percentage of nonzero coordinates that differ. |
| Chebychev| Chebychev distance (maximum coordinate difference).                         |

3.3. Selection of the best combination

Previously, there are 77 combinations for hierarchical clustering. Thus, the best combination of both measure and linkage will be evaluated based on few factors such as Malaysia geographical, monsoon limitation and the even segregation of the stations in the cluster. Result will be discussed in the next section.

4. Result and discussion

Annual summation of rainfall series for Peninsular Malaysia is used to obtain the optimal number of cluster. The clustering process using Hierarchical complete clustering and Euclidean distance has been done in this study. This method is used because it suits the data well and widely used by the researcher. Figure 3 shows the results of 8 validity indices used to evaluate the clustering solutions. The optimal number of cluster is indicated by a square symbol.
Table 3 summarized the results and it indicates that the best number of cluster is three since majority or 50% of the validity indices show the same result.

Table 3: Optimal number of cluster (Euclidean)
The final step is to allocate the rain gauge stations according to the appropriate cluster. By using hierarchical clustering, two best combinations of similarity measure and linkages were obtained. The first combination (i.e. Complete Cluster) is the combination of complete linkage and correlation similarity measures. Complete cluster segregated Peninsular Malaysia into three parts which are top, middle and bottom. In the top cluster, there are 21 stations from northern and eastern part of Peninsular Malaysia followed by middle cluster that consists of 19 stations mostly from western part. Bottom cluster comprised the southern part of Peninsular Malaysia with 12 stations altogether.

The second combination (i.e. Weighted cluster) is the combination of weighted linkage and correlation similarity measures. Peninsular Malaysia is separated into three parts and basically it has the same basis with Complete cluster except for the number of stations in each part. In the top part, there are 25 stations, for middle there are 16 stations and bottom part have 12 stations. Nevertheless, both figures have to eliminate 7 stations because of station mislocation. Figures 4(a) and Figure 4(b) illustrate the rain gauge stations over the Peninsular Malaysia according to the respective cluster. Red dots indicate the stations that are mislocated. In this study, we chose to remove these stations because special processes are needed to overcome mislocation problem.

| Method                          | Complete |
|---------------------------------|----------|
| a. Silhoutte index              | 2        |
| b. Davies-Bouldin               | 10       |
| c. Calinski-Harabasz            | 3        |
| d. Dunn Index                   | 2        |
| e. C Index                      | 3        |
| f. Krzanowski-Lai Index         | 3        |
| g. Hartigan Index               | 3        |
| h. Weighted Inter-Intra Index   | 4        |
5. Conclusion

A lot of studies with pertain to rainfall pattern of Malaysia have been done. It provides beneficial
information to various sectors including hydrological and water resources management. However, 59
rain gauge stations are too wide to be studied independently. Time and cost for the researcher will be
the greatest restrictions to complete the task. Thus, this study has been done to group the rain gauge
stations in Peninsular Malaysia into several clusters to reduce the masses of observation.

Three steps used in this study are the determination the optimal number of cluster, hierarchical
clustering and selection of the best combination. The optimal number of cluster for Peninsular
Malaysia is three. It represents Peninsular Malaysia as top, middle and bottom part, namely cluster A,
B and C. In this study, the combination between complete linkage and correlation similarity measures
is more suitable to fits the data set. This is based on the monsoon condition, the segregation of the
stations in each cluster and its simplicity. The border of the cluster for this method is
illustrated well because it equally divides Peninsular Malaysia as shown in Figure 5. The stations are
also randomly distributed within the cluster and fairly dispersed for each cluster. Further analysis of
this topic can be done and the seasonal climate pattern of Malaysia should be taken into account.

![Figure 5: Final Cluster Map of Peninsular Malaysia](image)

Acknowledgment

The authors are indebted to the staff of the Meteorological Department of Malaysia for providing the
daily rainfall data for this study. They also register their sincere appreciation to the reviewers for their
valuable suggestions and remarks which has improved the manuscript. This research is funded by
Universiti Teknologi MARA (600-RMI/ST/FRGS 5/3/Fst (192/2010)).
References

[1] Camerlengo A L, N Saadon, A S Aziz and R Othman 1998 Monthly Distribution of Precipitation in Peninsular Malaysia Pertanika Journal Science & Technology 6(1): 59 - 70 (1998).

[2] Leong W C 2007 Study On Malaysian Urban Rainfall-Runoff Characteristics: Case Study Of Sungai Kayu Ara, Damansara, SelangorMsc. Civil Engineering, Universiti Sains Malaysia.

[3] Suhaila J, Deni S M, W Zawiah and Jemain AA 2010 Trends In Peninsular Malaysia Rainfall Data During The Southwest Monsoon And Northeast Monsoon Seasons: 1975 - 2004 Sains Malaysia 39(4)(2010): 533 - 542.

[4] C L Wong, R Venneker, S Uhlenbrook, A B M Jamil and Y. Zhoul 2009 Variability of rainfall in Peninsular Malaysia Hydrology Earth System Sciences Discussion 6, 5471 - 5503.

[5] R Modarres 2006 Regional precipitation climates of Iran Journal of Hydrology (NZ) 45 (1): 13-27, 2006.

[6] Romero R, Summer G, Ramis C and Genoves A 1999 A classification of the atmospheric circulation patterns producing significant daily rainfall in the Spanish Mediterranean area International Journal of Climatol 19: 765 - 785

[7] Singh C V 1999 Principal components of monsoon rainfall in normal, flood and drought years over India International Journal of Climatology 19: 639-952.

[8] Stahl K., Demuth S 1999 Methods for regional classification of streamflow drought series: Cluster analysis Technical report to the ARIDE project, No.1

[9] Gottschalk L 1985 Hydrological regionalization of Sweden Hydrological Sciences Journal 30: 65-83.

[10] Guttman N B 1993 The use of L-Moments in the determination of regional precipitation climates Journal of Climate 6 (12): 2309-2325.

[11] Wiltshire S E 1986 Regional flood frequency analysis II: Multivariate classification of drainage basins in Britain Hydrological Sciences - Journal - 31, 3, 9/1986.

[12] Donald H. Burn 1989 Cluster Analysis As Applied To Regional Flood Frequency Journals/ Water Resources Planningand Management, Vol. 115, No. 5, September, 1989.

[13] Ramos M C 2001 Divisive and hierarchical clustering techniques to analyze variability of rainfall distribution patterns in a Mediterranean region Journal of Hydrology 57: 123-138.

[14] Richard A Johnson and Dean W Wichern 1982 Applied Multivariate Statistical Analysis © 1982 by Prentice-Hall, Inc.,Eaglewood Cliffs, NJ 07632.

[15] Maria Halkidi, Yannis Batistakis And Michalis Vazirgiannis 2001 On Clustering Validation Techniques Journal of Intelligent Information Systems, 17:2/3, 107–145, 2001

[16] Kaufman L, Rousseeuw P J 1990 Finding groups in data: An introduction to cluster analysis Wiley, New York, 344 pp.

[17] Deni S M, A A Jemain and Ibrahim K 2009 The best probability models for dry and wet spells in Peninsular Malaysia during monsoon seasons International Journal Of Climatology.

[18] N H Ahmad and Deni S M 2012 Homogeneity Test on Daily Rainfall Series for Malaysia Proceeding for 1” ISM International Statistical Conference 2012.

[19] F Yunus Delineation of Climate Divisions for Peninsular Malaysia Meteorological Officer (Report) of Meteorological Department of Malaysia