MALAYSIAN METEOROLOGICAL DEPARTMENT
MINISTRY OF SCIENCE, TECHNOLOGY AND INNOVATION

MALAYSIAN METEOROLOGICAL DEPARTMENT (MMD)
MINISTRY OF SCIENCE, TECHNOLOGY AND INNOVATION (MOSTI)

Research Publication No. 5/2017

Pengelompokan Corak Taburan Hujan dengan Kaedah Pengelompokan Siri Masa

Sharifah Faridah S.M.

RESEARCH PUBLICATION NO. 5/2017

PENGELOMPOKAN CORAK TABURAN HUJAN DENGAN KAEDAH PENGELOMPOKAN SIRI MASA

By Sharifah Faridah S. M.

Perpustakaan Negara Malaysia Cataloguing in Publication Data

Published and printed by

Malaysian Meteorological Department
Jalan Sultan
46667 PETALING JAYA
Selangor Darul Ehsan
Malaysia

Kandungan

Pengelompokan Corak Taburan Hujan Dengan Kaedah Pengelompokan Siri Masa (Clustering Of Rainfall Distribution Patterns Using Time Series Clustering Method)

Sharifah Faridah S. M.

ABSTRAK

Pengelompokan siri masa adalah salah satu konsep penting dalam perlombongan data untuk mengenal pasti kelompok set objek yang mana kelasnya tidak diketahui. Siri hujan dari tahun 1970 hingga 2014 bagi 12 stesen pencerapan meteorologi yang homogen telah digunakan dalam kajian pengelompokan corak taburan hujan di Semenanjung Malaysia. Salah satu komponen utama dalam menentukan hasil analisis kelompok yang bermakna ialah dalam menentukan sukatan ketaksamaan yang tepat dan sesuai. Sebanyak empat kaedah sukatan ketaksamaan telah dikaji ketepatan dan kesesuaian terhadap data hujan iaitu jarak Euclidean (ED), complexity-invariant (CID), jarak berasaskan korelasi (COR) dan jarak berasaskan periodogram bersepadu (IP). Kaedah purata lebar rupa bentuk (ASW) telah digunakan dalam menentukan bilangan kelompok yang optimum bagi siri masa hujan. Dengan menggunakan kaedah pengelompokan berhierarki Ward, kajian mendapati sebanyak empat rantau dapat dibahagikan mengikut zon klimatologi yang homogen untuk data siri masa hujan Semenanjung Malaysia secara keseluruhan dan musiman; Monsun Timur Laut dan Monsun Barat Daya.

Kata Kunci: pengelompokan siri masa; sukatan ketaksamaan; analisis kelompok

ABSTRACT

Time series clustering is one of the important concepts in data mining to identify the set of objects whose class is unknown. Rainfall series from 1970 to 2014 from 12 homogenous meteorological observation stations were used in the study of clustering rainfall patterns in Peninsular Malaysia. One of the key components in determining the significant result of a cluster analysis is in choosing the right and appropriate dissimilarity measures. Four methods of dissimilarity measures were examined for their accuracy and suitability to the rainfall data including the Euclidean distance (ED), complexity-invariant (CID), correlation-based distance (COR) and integrated periodogram-based distance (IP). The average silhouette width (ASW) was used to determine the optimal number of groups for a time series of rainfall. Using Ward's hierarchical clustering method, the study found that the four regions can be divided into homogeneous climatological zones for time series of rainfall in Peninsular Malaysia as a whole and seasonal; Northeast Monsoon and Southwest Monsoon.

Keywords: time series clustering; dissimilarity measures; cluster analysis

1.0 PENDAHULUAN

Ketepatan dalam peramalan cuaca memberikan kesan dan impak yang tinggi ke atas aktiviti sosio-ekonomi dan pembangunan sesebuah negara. Ianya dapat dilihat bila mana maklumat dan laporan cuaca ini digunakan untuk membuat sesuatu keputusan penting dalam pelbagai sektor antaranya sektor pengurusan bencana, pengurusan sumber air negara, pertanian, perindustrian dan pelancongan. Sebagai contoh, laporan dari Perancangan Strategik *World Meteorological Organization* (WMO) untuk 2012-2015 melaporkan bahawa anggaran keuntungan, dari segi ekonomi, kepada sektor pertanian, mencecah antara USD 450-550 juta setahun, hasil daripada ketepatan maklumat dan ramalan cuaca berkaitan dengan fenomena El-Nino (WMO 2011). Hal ini boleh terlaksana dengan melakukan pelbagai kajian bagi meningkatkan prestasi ramalan cuaca dan iklim. Antara kajian yang dilakukan adalah dengan meramalkan kejadian cuaca luar jangka seperti kejadian taufan, ribut dan banjir menggunakan kaedah perlombongan data (Bartok et al. 2010; Kohail & El-Halees 2011; Mohammadi et al. 2006).

Teknik pengelompokan merupakan salah satu kaedah perlombongan data yang efektif dalam memberikan maklumat berguna bagi pelbagai cabang penyelidikan saintifik. Ini jelas dapat dilihat melalui penerbitan buku-buku teks dan lain-lain terbitan berkaitan pengelompokan ke atas data-data saintifik seperti taksonomi, pertanian, penderiaan jauh dan kawalan proses (Kavitha & Punithavalli 2010). Kaedah pengelompokan ini juga digunakan dengan meluas dalam bidang hidrologi, klimatologi dan meteorologi untuk menentukan dan mengelaskan corak taburan hujan (Ahmad et al. 2013; Munoz-Dias & Rodrigo 2004; Soltani & Modarres 2006). Teknik pengelompokan siri masa juga turut digunakan dalam kajian ramalan seperti ramalan jualan yang dijalankan oleh Sanwlni dan Vijayalakshmi (2013).

Dalam kajian ini, maklumat berguna daripada rekod siri masa data hujan harian di 12 stesen pencerapan meteorologi Semenanjung Malaysia dikaji secara terperinci menggunakan teknik pengelompokan. Tempoh data yang dianalisis adalah 45 tahun (1970-2014). Tempoh yang dicadangkan oleh WMO untuk analisis data hujan adalah 30 tahun (Ahmad et al. 2013; Munoz-Dias & Rodrigo 2004). Pemprosesan data bermula dengan analisis kehomogenan, analisis data hilang dan diikuti penilaian ke atas sukatan ketaksamaan yang digunakan dalam teknik pengelompokan.

2.0 OBJEKTIF

1. Mengkaji kaedah pengelompokan berhierarki yang paling sesuai dalam menentukan kelompok siri masa bagi kebarangkalian berlakunya hujan harian

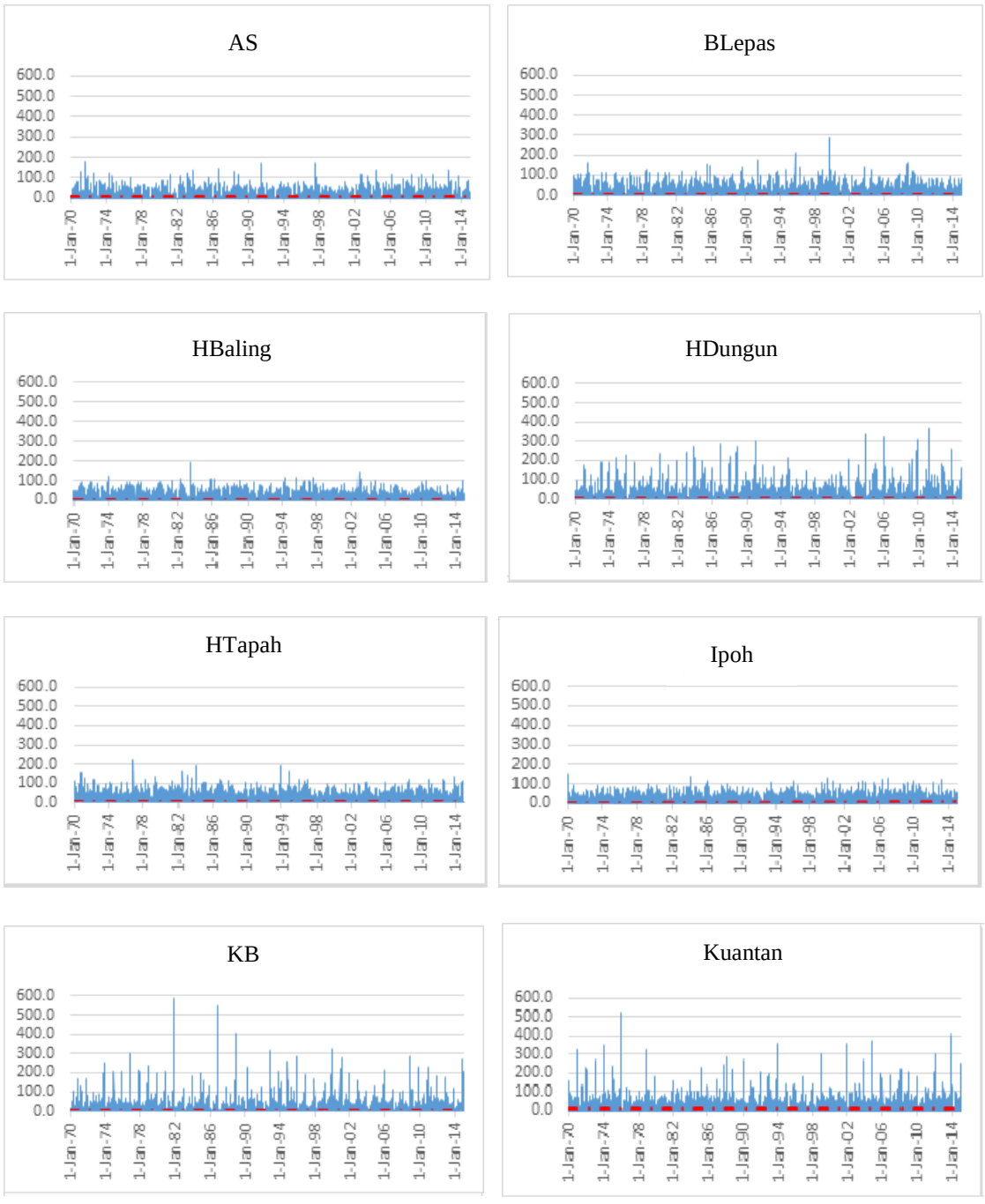
2. Mengkaji sukatan ketaksamaan dalam analisis pengelompokan siri masa yang paling sesuai bagi data hujan di Semenanjung Malaysia

3. Menguji keberkesanan kaedah pengelompokan yang digunakan dan membandingkan hasil kelompok yang diperoleh dengan kelompok hujan sedia ada

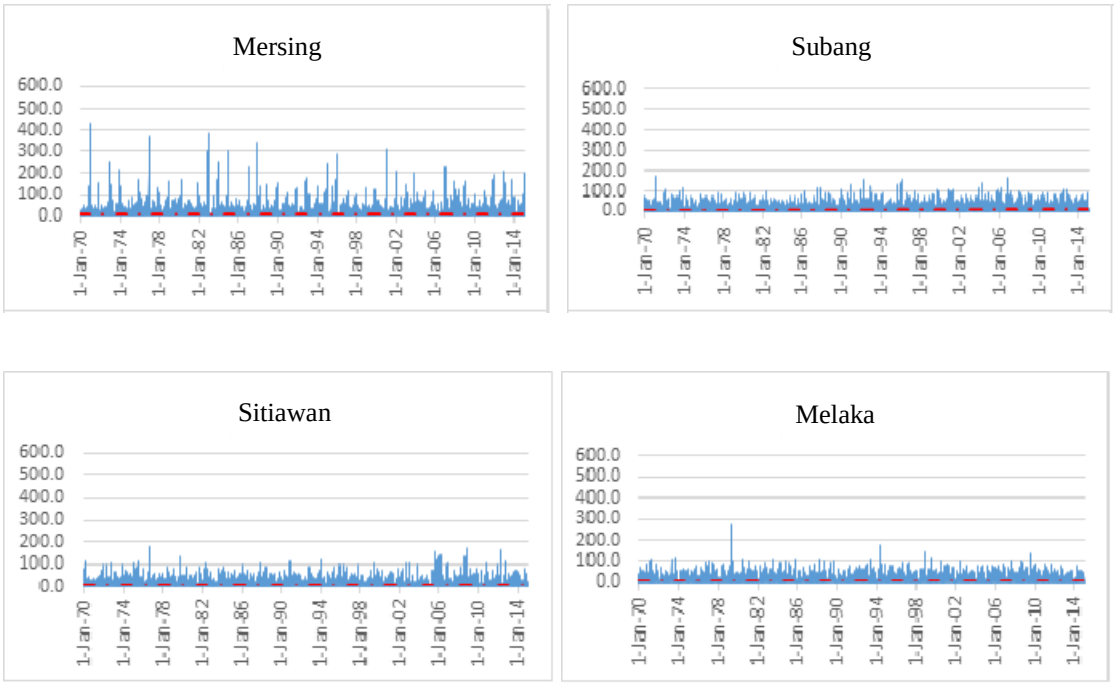
3.0 METODOLOGI

Dalam kajian ini, data hujan harian dari 12 stesen pencerapan data meteorologi, Jabatan Meteorologi Malaysia (JMM) telah digunakan. Tempoh data adalah dari tahun 1970 sehingga 2014. Taburan hujan bagi 12 stesen dipaparkan dalam Rajah 3.1, lokasi stesen dipetakan dalam Rajah 3.2 dan maklumat data serta stesen diberikan dalam Jadual 3.1.

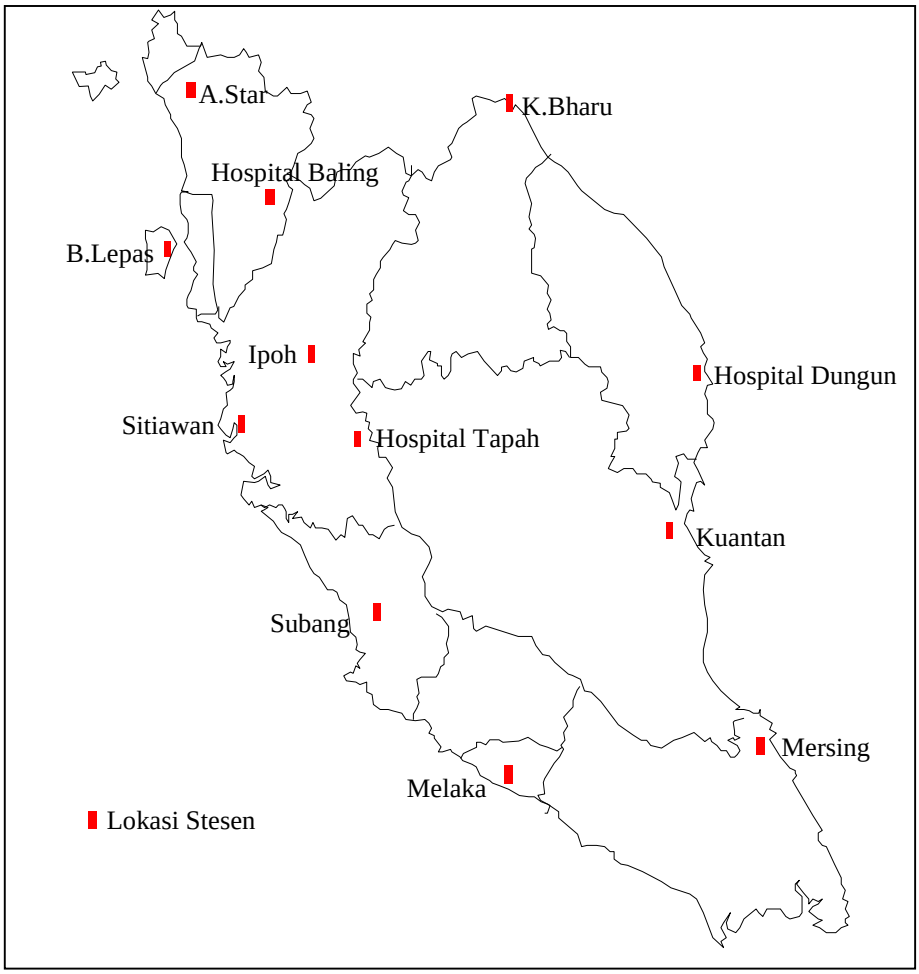
Sebanyak 12 stesen pencerapan data meteorologi yang dikendalikan oleh JMM (Rajah 3.1) telah digunakan dalam kajian ini. Stesen-stesen dipilih meliputi tiga zon di Semenanjung iaitu bahagian barat laut, barat dan timur. Stesen di bahagian barat laut terdiri daripada Stesen Meteorologi Alor Setar (Kedah), Hospital Baling (Kedah) dan Bayan Lepas (Pulau Pinang). Stesen di bahagian barat pula terdiri daripada Stesen Meteorologi Ipoh (Perak), Sitiawan (Perak), Hospital Tapah (Perak), Subang (Selangor) dan Melaka (Melaka). Manakala stesen-stesen di bahagian timur adalah Stesen Meteorologi Kota Bharu (Kelantan), Hospital Dungun (Terengganu), Kuantan (Pahang) dan Mersing (Johor). Pemilihan stesen-stesen tersebut adalah berdasarkan kepada kedudukan geografi dan juga ketersediaan data di kawasan tersebut (Jadual 3.1). Diperhatikan dalam Rajah 3.2, tiada stesen dipilih di kawasan tengah berikutan daripada ketiadaan stesen yang merekodkan data lengkap untuk tempoh 45 tahun.



Rajah 3.1 Taburan data hujan harian (mm) dari tahun 1970 hingga 2014 di stesen-stesen meteorologi pilihan



Rajah 3.1 (Sambungan)



Rajah 3.2 Lokasi stesen-stesen pencerapan data hujan

Jadual 3.1 Butiran nombor stesen, lokasi dan peratusan data hilang untuk setiap stesen

Nama Stesen	Kod Stesen	Latitud (°N)	Longitud (°E)	Ketinggian dari Aras Laut (MSL) (m)	Data Hilang (%)
ALOR SETAR	48603	6.2	100.4	3.9	-
BAYAN LEPAS	48601	5.3	100.2667	2.5	-
KOTA BHARU	48615	6.1667	102.3	4.4	-
HOSPITAL DUNGUN	49476	4.7667	103.4167	3	2.46
KUANTAN	48657	3.7667	103.2167	15.2	-
MERSING	48674	2.45	103.8333	43.6	-
SUBANG	48647	3.1333	101.55	16.6	-
MALACCA	48665	2.2667	102.25	8.5	-
HOSPITAL BALING	41545	5.6833	100.9167	52	0.42
IPOH	48625	4.5667	101.1	40.1	-
HOSPITAL TAPAH	43421	4.2	101.2667	35	0.27
SITIAWAN	48620	4.2167	100.7	6.8	-

3.1 UJIAN KEHOMOGENAN

Tahap kualiti data hujan di stesen pencerapan perlu dipastikan berada dalam keadaan yang baik supaya hasil kajian mempunyai tahap kebolehpercayaan yang tinggi. Data-data hujan yang dicerap mempunyai tahap kebergantungan yang tinggi terhadap faktor-faktor luaran seperti lokasi stesen, alat pencerapan dan kaedah bacaan data diambil. Disebabkan itu, data-data meteorologi ini perlu diuji secara statistik (Mahmud Firat et al. 2010).

Ianya boleh ditafsirkan bahawa struktur semulajadi nilai pemerhatian ke atas data yang dicerap tidak merosot apabila siri masa hujan mempunyai struktur yang homogen (Mahmud Firat et al. 2010). Melalui ujian kehomogenan, rekod data yang menunjukkan ketakhomogenan dapat dikesan dan diubahsuai ataupun dibuang.

Sehubungan dengan itu, ujian Bartlett telah dijalankan untuk menguji tahap kehomogenan data yang digunakan dalam kajian ini. Hasil ujian dengan nilai –p yang lebih rendah daripada tahap keertian –$\alpha = 0.05$ menunjukkan ketakhomogenan dan data tidak mewakili data iklim sebenar. Ujian Bartlett bagi menguji kehomogenan data, telah digunakan oleh Estaban-Parra et al. (1998) dalam kajiannya terhadap corak hujan musiman di Sepanyol (Munoz-Diaz & Rodrigo 2004).

3.2 ANALISIS DATA HILANG

Analisis data hilang adalah merupakan antara langkah utama sebelum proses analisis selanjutnya dapat dijalankan. Sebagaimana data-data siri masa yang lain, set siri masa data hujan dalam kajian ini turut merekodkan data hilang seperti yang dipaparkan dalam Jadual 3.1.

Stesen-stesen yang merekodkan data hilang telah dikenal pasti iaitu Stesen Meteorologi Hospital Dungun (2.46%), Hospital Baling (0.42%) dan Hospital Tapah (0.27%) dan didapati peratusan data hilang adalah kecil. Menurut Johnson (2003), apa-apa kaedah penggantian atau interpolasi data boleh dijalankan dengan hasil keputusan yang baik bagi rekod data hilang kurang daripada 5% (Bhotale & Katpatal 2014). Dengan itu, adalah tidak praktikal untuk membuang data-data ini, sebaliknya kaedah penggantian data hilang dengan menggunakan nilai purata mengikut kumpulan kecil telah dijalankan dalam kajian ini. Kaedah ini menggantikan nilai-nilai data hilang dengan nilai purata atribut yang diketahui mengikut kumpulan yang dibahagikan (Kaiser 2014). Menurut Acock (2005), penggunaan teknik

penggantian data hilang menggunakan nilai purata mengikut kumpulan yang dibahagikan dapat memberikan anggaran yang lebih baik dan dapat memelihara lebih varian berbanding menggunakan nilai purata keseluruhan.

3.3 PENGELOMPOKAN HIERARKI KAEDAH WARD

Kaedah Ward yang digunakan dalam pengelompokan berhierarki ini meminimakan kehilangan maklumat hasil daripada gabungan kelompok-kelompok. Pada setiap peringkat, gabungan ke atas setiap pasangan kelompok yang mungkin, dipertimbangkan dan dua kelompok dengan hasil gabungan yang memberikan nilai peningkatan ralat jumlah kuasa dua (ESS) yang kecil akan digabungkan. Akhirnya, kesemua kelompok digabungkan menjadi satu kelompok yang besar dengan nilai ESS dikira menggunakan formula seperti berikut:

$$\text{ESS} = \sum_{j=1}^{N} (x_j - \bar{x})' (x_j - \bar{x}) \qquad \ldots(3.1)$$

Dengan,

x_j Pengukuran multivariat berkaitan item j

$\bar{x}$ Nilai purata kesemua item

3.4 SUKATAN KETAKSAMAAN

Sebelum menggunakan algoritma pengelompokan siri masa, sukatan kesamaan dan ketaksamaan berangka bagi mencirikan hubungan antara data perlu dikenal pasti. Sebanyak empat kaedah sukatan ketaksamaan bagi tujuan menentukan sukatan ketaksamaan yang paling sesuai untuk data hujan di Semenanjung Malaysia digunakan dalam kajian ini. Sukatan ketaksamaan yang digunakan ialah Jarak *Euclidean*, *Complexity-invariant*, Jarak Berasaskan Korelasi dan Jarak Berasaskan Periodogram Bersepadu.

3.4.1 Jarak *Euclidean*

Jarak *Euclidean* (ED) merupakan sukatan ketaksamaan yang paling mudah untuk siri masa (Faloutsos et al. 1994). Sukatan ED ini merupakan sukatan yang paling biasa digunakan walaupun wujudnya banyak sukatan lain (Munoz-Dias & Rodrigo 2004). Tambahan pula, sukatan ketaksamaan menggunakan ED ini setanding dengan pendekatan lain yang lebih

kompleks terutamanya apabila melibatkan saiz set latihan atau pangkalan data yang agak besar (Ding et al. 2008). Formula ED adalah seperti berikut:

$$\text{ED}\ (\boldsymbol{X}_T\ ,\ \boldsymbol{Y}_T) = (\sum_{t=1}^{T}(X_T - Y_T\ )^2\ )^{1/2} \qquad \ldots(3.2)$$

Dengan,

$\boldsymbol{X}_T$ = Siri masa $(X_1\ , X_2, X_3\ \ldots X_T)'$

$\boldsymbol{Y}_T$ = Siri masa $(Y_1\ , Y_2, Y_3\ \ldots Y_T)'$

3.4.2 C*omplexity-invariant*

Batista et al. (2011) telah memperkenalkan sukatan ketaksamaan *complexity-invariant* (CID) untuk siri masa. Kaedah CID ini adalah kaedah tambahan kepada ED. CID merupakan faktor perbezaan kompleks antara dua siri masa. Ia dikira sebagai nisbah siri yang lebih kompleks terhadap siri kurang kompleks. Faktor pembetulan kompleks menghampiri nilai satu apabila kedua-dua siri masa mempunyai kerumitan yang sama (sama ada keduanya adalah sangat kompleks atau sebaliknya). Apabila kedua-dua siri masa mempunyai kerumitan yang tidak sama, faktor pembetulan kompleks akan memberikan nilai melebihi satu. Formula CID seperti berikut:

$$\text{CID}\ (\boldsymbol{X}_T\ ,\ \boldsymbol{Y}_T) =\ \text{ED}\ (\boldsymbol{X}_T\ ,\ \boldsymbol{Y}_T)\ \ \text{x CF}\ (\boldsymbol{X}_T\ ,\ \boldsymbol{Y}_T) \qquad \ldots(3.3)$$

di mana CF adalah faktor pembetulan kompleks,

$$\text{CF}\ (\boldsymbol{X}_T\ ,\ \boldsymbol{Y}_T)\ =\ \frac{max(\text{CE}\ (X_T\ ),\ CE\ (Y_T))}{min(\text{CE}\ (X_T\ ),\ CE\ (Y_T))} \qquad \ldots(3.4)$$

Dengan CE($\boldsymbol{X}_T$), adalah penganggar kompleks,

$$\text{CE}(\boldsymbol{X}_T) = \sqrt{\sum_{t=1}^{T-1}(X_t - X_{t+1}\ )^2} \qquad \ldots(3.5)$$

3.4.3 Jarak berasaskan korelasi

(*Correlation-based distances*)

Sukatan ketaksamaan jarak berasaskan korelasi yang dipertimbangkan dalam kajian ini ialah faktor korelasi Pearson (COR). Semakin tinggi nilai korelasi bermaksud semakin dekat jarak (Falcone & Albuquerque 2004). Formula untuk kiraan korelasi Pearson adalah seperti berikut:

$$\text{COR}\ (\boldsymbol{X}_T\ ,\ \boldsymbol{Y}_T)\ =\ \frac{\sum_{t=1}^{T}(X_t - \bar{X}_T)\ (Y_t - \bar{Y}_T)}{\sqrt{\sum_{t=1}^{T-1}(X_T - \bar{X}_T)^2}\ \sqrt{\sum_{t=1}^{T-1}(Y_T - \bar{Y}_T)^2}} \qquad \ldots(3.6)$$

Dengan,

$\bar{\boldsymbol{X}}_T$ = Nilai purata siri masa $\boldsymbol{X}_T$

$\bar{\boldsymbol{Y}}_T$ = Nilai purata siri masa $\boldsymbol{Y}_T$

3.4.4 Jarak berasaskan periodogram bersepadu

(*Periodograms-based distances: Integrated Periodogram*)

Kaedah periodogram digunakan untuk mengenal pasti tempoh dominan atau pun frekuensi sesuatu siri masa. Ianya boleh menjadi satu alat yang berguna untuk mengenal pasti tingkah laku kitaran yang dominan dalam siri masa. Ia juga merupakan salah satu cara yang boleh digunakan untuk menganalisis data berkala dengan menghuraikan data-data tersebut ke dalam bentuk gelombang pada frekuensi yang berbeza-beza. Prosedur ini berguna untuk menetapkan keadaan rawak dan musiman ke atas data berkala dan juga berguna untuk mengenali adanya autokorelasi positif dan negatif.

Casado de Lucas (2010) mempertimbangkan pengiraan jarak ke atas periodogram secara kumulatif dan memperkenalkan pengiraan sukatan jarak periodogram bersepadu (IP) (Montero & Vilar 2014). Sukatan jarak IP ini mengira perbezaan jarak antara dua siri masa dari segi jarak antara periodogram bersepadu mereka. Langkah – langkah pengiraan adalah seperti berikut:

$$\text{IP}\ (\boldsymbol{X}_T\ ,\ \boldsymbol{Y}_T) = \int_{-\pi}^{\pi}\left|F_{X_T}\ (\lambda) -\ F_{Y_T}\ (\lambda)\right|\ d\lambda \quad ,\lambda\ \in [-\pi, \pi] \qquad \ldots(3.7)$$

$$F_{X_T}(\lambda_j) = C_{X_T}^{-1} \sum_{i=1}^{j} I_{X_T}(\lambda_i) \quad , C_{X_T} = \sum_i I_X(\lambda_i) \qquad \ldots (3.8)$$

$$F_{Y_T}(\lambda_j) = C_{Y_T}^{-1} \sum_{i=1}^{j} I_{Y_T}(\lambda_i) \quad , C_{Y_T} = \sum_i I_Y(\lambda_i) \qquad \ldots (3.9)$$

$$I_{X_T}(\lambda_k) = T^{-1} \left| \sum_{t=1}^{T} X_T \; e^{-i\lambda_k t} \right|^2 \qquad \ldots (3.10)$$

$$I_{Y_T}(\lambda_k) = T^{-1} \left| \sum_{t=1}^{T} Y_T \; e^{-i\lambda_k t} \right|^2 \qquad \ldots (3.11)$$

$$\lambda_k = \frac{2\pi k}{T} \quad , k = 1, \ldots, n \qquad \ldots (3.12)$$

$$n = \left[\frac{T-1}{2} \right] \qquad \ldots (3.13)$$

Dengan,

T = Panjang vektor, $T \geq 1$

$I_{X_T}(\lambda_k)$ = Periodogram bagi $\boldsymbol{X}_T$

$I_{Y_T}(\lambda_k)$ = Periodogram bagi $\boldsymbol{Y}_T$

3.5 Purata lebar rupa bentuk
(Average Silhouette Width)

Menentukan bilangan kelompok yang optimum dalam satu set data, *k,* merupakan perkara asas dalam proses pengelompokan. Terdapat pelbagai kaedah yang boleh digunakan dalam menentukan nilai *k* ini. Dalam kajian ini, nilai purata lebar rupa bentuk (ASW) akan dikira dan digunakan. Kaedah ASW ini diperkenalkan oleh Rousseeuw pada tahun 1986 dalam kajiannya berkaitan dengan analisis pengelompokan (Potocnik et al. 2011).

Pada peringkat awal, purata jarak setiap individu dalam kelompok yang sama dikira. Ahli yang memberikan bacaan jarak yang rendah menunjukkan perbezaan antara individu dalam kelompok adalah minima dan sesuai ditempatkan dalam kelompok yang ditetapkan. Akhirnya, purata jarak setiap individu yang diperoleh akan dibandingkan dengan purata jarak ahli-ahli kelompok jiran. Nisbah perbezaan yang diperoleh dari titik ketaksamaan ahli dalam kelompok yang sama terhadap kelompok jiran terhampir dikenali sebagai nilai rupa bentuk. Nilai rupa bentuk keseluruhan kelompok dikira sebagai purata rupa bentuk setiap ahli. Ini mengukur tahap persamaan ahli kelompok. Nilai ASW yang diperoleh akhirnya, dapat menentukan bilangan kelompok optimum dalam set data, k.

ASW ini membantu dalam penilaian terhadap kesahihan proses pengelompokan yang dilakukan dan seterusnya memilih nombor yang optimum bagi kelompok dalam set data yang dianalisis. Menurut Potocnik et al. 2011, penilaian kelompok menggunakan kaedah ASW dapat megesahkan kualiti yang baik terhadap hasil kelompok yang membawa maksud kelompok yang dicadangkan itu adalah yang bermakna.

4.0 HASIL DAN PERBINCANGAN

Sebanyak empat teknik ukuran dijalankan bagi menentukan sukatan ketaksamaan yang paling sesuai dalam analisis pengelompokan siri masa data hujan di Semenanjung Malaysia. Hasil pengelompokan kaedah Ward dengan ukuran sukatan ketaksamaan berbeza dibandingkan dan ukuran terbaik dikenal pasti. Didapati bilangan kelompok yang paling optimum ke atas set data kajian ini adalah $k = 4$. Nilai k ini diperoleh menggunakan teknik ASW dan diguna pakai dalam proses analisis pengelompokan yang dijalankan.

4.1 PENILAIAN KEHOMOGENAN

Hasil daripada ujian Bartlett, didapati kesemua rekod siri masa data hujan di 12 stesen pencerapan tidak mempunyai masalah ketakhomogenan dan kualiti data berada pada tahap yang baik. Ini bermaksud, rekod data 12 stesen pencerapan yang digunakan dalam kajian, bebas daripada faktor perubahan bukan-iklim. Tambahan pula, sembilan daripada 12 stesen yang digunakan dalam kajian ini juga telah disahkan homogen seperti yang dilaporkan oleh Ahmad dan Deni (2013) dalam kajian kehomogenan terhadap siri masa data hujan harian di Malaysia. Ringkasan keputusan ujian kehomogenan dipaparkan dalam Jadual 4.1.

Jadual 4.1 Keputusan ujian Bartlett rekod siri masa hujan tahunan (1970-2014) mengikut zon.

Stesen	k-kuasa dua Bartlett	Darjah Kebebasan	Nilai-p
Zon Timur:			
1. Kota Bharu 2. Hospital Dungun 3. Kuantan 4. Mersing	1.2171	3	0.7489
Zon Barat Laut:			
1. Alor Setar 2. Hospital Baling 3. Bayan Lepas	2.5603	2	0.278
Zon Barat I:			
1. Ipoh 2. Hospital Tapah 3. Subang	0.6254	2	0.7315
Zon Barat II:			
1. Sitiawan 2. Melaka	0.0883	1	0.7663

4.2 PENILAIAN SUKATAN KETAKSAMAAN

Penilaian ke atas empat sukatan ketaksamaan dilakukan dengan nilai paling hampir dengan 1 adalah sukatan ketaksamaan yang paling sesuai untuk data siri masa kajian ini. Sebanyak tiga set data siri masa dianalisis yang melibatkan data hujan secara keseluruhan dan secara musiman. Bagi analisis musiman, dua monsun utama negara iaitu Monsun Timur Laut (November – Februari) dan Monsun Barat Daya (Mei – Ogos) dianalisis. Penilaian dilakukan ke atas setiap set data siri masa yang melibatkan 12 stesen di Semenanjung Malaysia. Bilangan kelompok yang digunakan adalah nilai optimum yang diperoleh dari kaedah ASW yang dijalankan iaitu $k = 4$. Hasil penilaian yang dijalankan ke atas siri masa diringkaskan dalam Jadual 4.2.

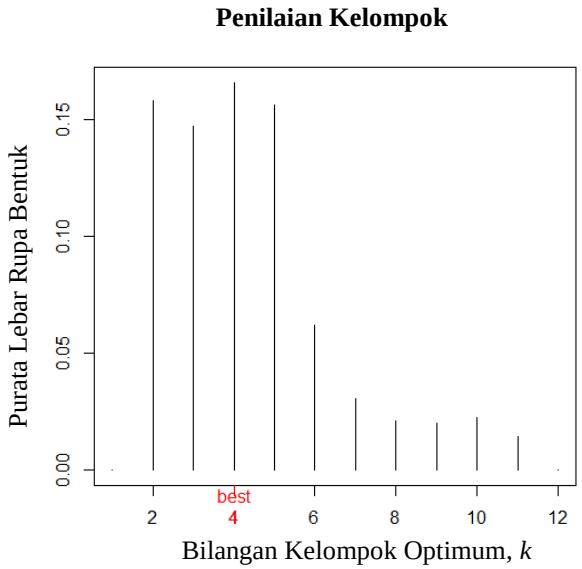
Jadual 4.2 Hasil penilaian sukatan ketaksamaan ke atas analisis siri masa data hujan keseluruhan, Monsun Timur Laut (MTL) dan Monsun Barat Daya (MBD).

Jarak Sukatan Ketaksamaan	Siri Masa		
	Keseluruhan	MTL	MBD
ED	0.4977	0.4977	0.5583
CID	0.5857	0.6167	0.425
IP	0.7292	0.6167	0.6167
COR	0.6792	0.5778	0.6786

Berdasarkan hasil penilaian Jadual 4.2, didapati sukatan jarak ketaksamaan IP dapat menghasilkan analisis kelompok yang terbaik berbanding sukatan yang lain ke atas siri masa bagi data keseluruhan dan musiman tempoh MTL. Nilai menghampiri satu yang diperoleh menunjukkan persetujuan yang ketara antara pembahagian kelompok data sebenar dan model. Sukatan ketaksamaan CID juga diperhatikan dapat menyumbang kepada hasil analisis kelompok terbaik bagi tempoh MTL. Berbeza bagi tempoh MDB, didapati sukatan ketaksamaan COR dapat memberikan hasil analisis kelompok terbaik dalam kajian ini. Manakala penilaian ke atas sukatan ketaksamaan ED didapati memberikan hasil yang kurang memuaskan untuk ketiga-tiga set siri masa.

4.3 ANALISIS DENDOGRAM: DATA KESELURUHAN

Dendogram hasil pengelompokan siri masa kaedah Ward dengan bilangan kelompok optimum k = 4 (Rajah 4.1) menggunakan empat teknik sukatan ketaksamaan dipaparkan dalam Rajah 4.2. Analisis ini menggunakan set data hujan harian keseluruhan (total) bagi tempoh 1970 – 2014 dari 12 stesen pencerapan.



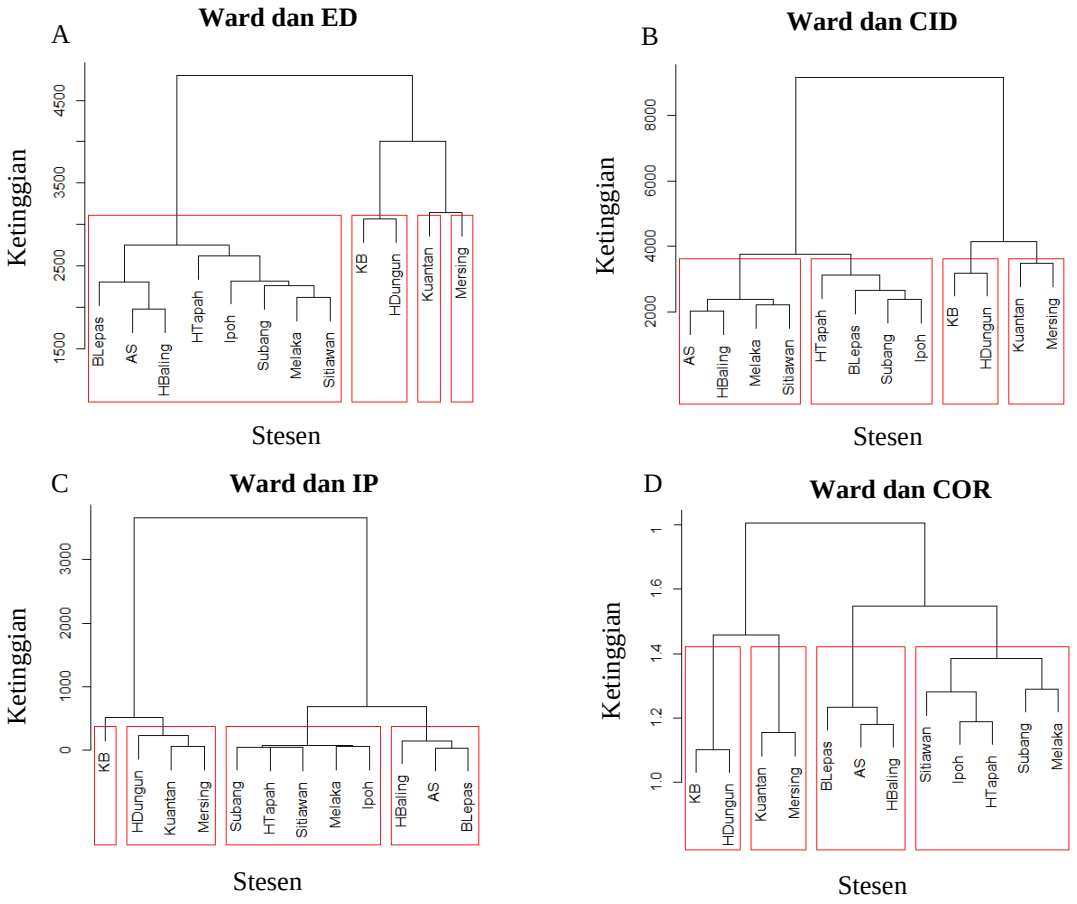
Rajah 4.1 Bilangan kelompok optimum k dengan kaedah ASW.

Kaedah Ward akan mengelompokkan kesemua stesen menjadi satu kelompok besar seperti yang dipaparkan dalam Rajah 4.2. Paksi –y yang dilabel sebagai ketinggian merujuk kepada perbezaan ataupun ketaksamaan antara setiap kelompok. Lebih panjang garis vertikal atau menegak yang dipaparkan menunjukkan perbezaan yang lebih besar di antara kelompok.

Berdasarkan dendogram dalam Rajah 4.2 didapati keputusan analisis kelompok yang hampir sama diperoleh untuk dua ukuran jarak ketaksamaan IP dan COR. Kedua-duanya mengelompokkan dua set kelompok stesen bersama-sama iaitu {Hospital Baling, Alor Setar, Bayan Lepas} dan {Tapah, Sitiawan, Ipoh, Subang, Melaka}. Ini menunjukkan bahawa stesen-stesen yang dikelompokkan dalam kelompok yang sama adalah agak kompleks untuk mengetahui perbezaan yang besar di kalangan stesen tersebut. Seterusnya IP mengelompokkan {Kuantan, Mersing, Hospital Dungun} bersama-sama, manakala COR mengelompokkan {Kuantan, Mersing} dan {Kota Bharu, Hospital Dungun}. Kota Bharu dikenal pasti sebagai objek terpencil menggunakan ukuran IP.

CID dan ED turut mengelompokkan {Kota Bharu, Hospital Dungun} bersama-sama. Stesen selebihnya dengan ukuran CID, memberikan hasil kelompok yang berikut; {Alor Setar, Hospital Baling, Sitiawan, Melaka} dan {Bayan Lepas, Hospital Tapah, Ipoh, Subang}. ED pula mengelompokkan stesen-stesen {Alor Setar, Hospital Baling, Bayan Lepas, Hospital Tapah, Ipoh, Sitiawan, Subang, Melaka} dalam satu kelompok yang sama. Kuantan dan Mersing masing-masing dikenal pasti sebagai objek terpencil menggunakan ED.

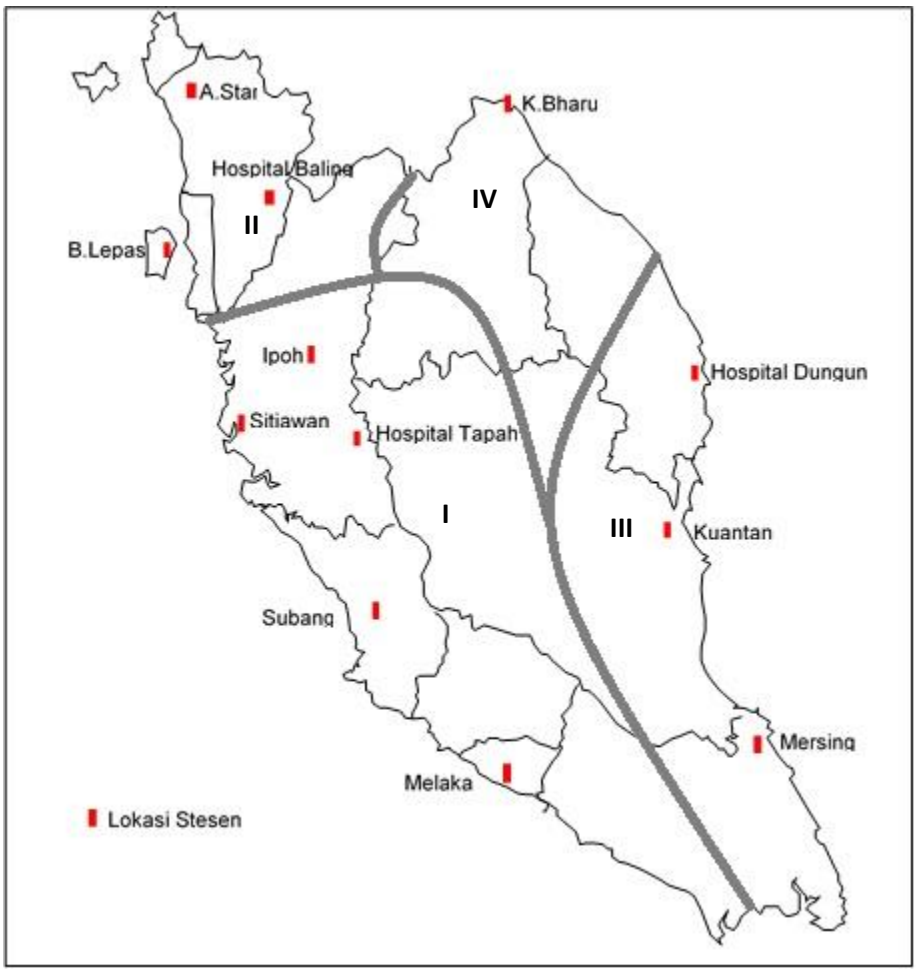
Didapati faktor geografi dan kedudukan stesen secara semulajadi sangat mempengaruhi kelompok-kelompok yang terhasil. Ringkasan keputusan analisis dendogram bagi setiap ukuran jarak ketaksamaan mengikut kelompok dalam peratusan diberikan dalam Jadual 4.3. Manakala, pembahagian sempadan hasil pengelompokan menggunakan sukatan ketaksamaan IP, iaitu hasil kelompok terbaik (rujuk Jadual 4.2) untuk data keseluruhan, diilustrasikan dalam Rajah 4.3.



Rajah 4.2. Dendogram untuk 12 stesen bagi analisis data siri masa tempoh 1970 – 2014 dengan sukatan ketaksamaan A) Jarak *Euclidean* B) *Complexity-invariant* C) *Integrated Periodogram-based* D) *Correlation-based*.

Jadual 4.3 Peratusan (%) bilangan stesen mengikut kelompok dengan sukatan ketaksamaan berbeza bagi analisis data keseluruhan.

Kelompok	**Jarak Sukatan Ketaksamaan**			
	ED	**CID**	**IP**	**COR**
#1	66.70%	33.30%	41.70%	41.70%
#2	16.70%	33.30%	25.00%	25.00%
#3	8.30%	16.70%	25.00%	16.70%
#4	8.30%	16.70%	8.30%	16.70%



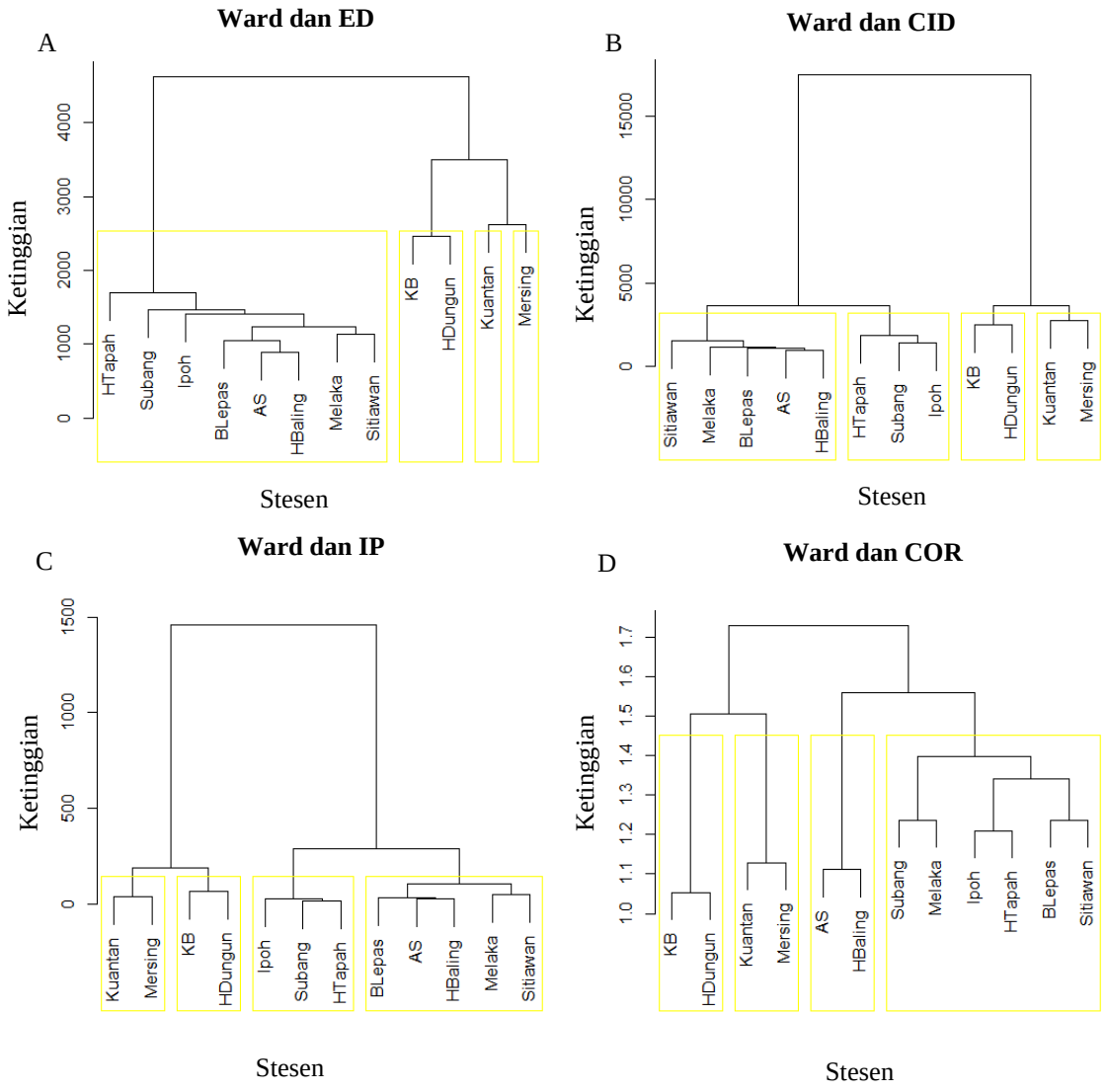
Rajah 4.3 Peta pengelompokan data hujan Semenanjung Malaysia menggunakan sukatan IP terhadap data keseluruhan.

4.4 ANALISIS DENDOGRAM: DATA MUSIMAN MTL

Berdasarkan dendogram dalam Rajah 4.4 didapati analisis kelompok menggunakan sukatan ketaksamaan CID dan IP memberikan hasil keputusan yang sama. Kelompok yang dikenal pasti ialah {Kota Bharu, Hospital Dungun}, {Kuantan, Mersing}, {Ipoh, Hospital Tapah, Subang} dan {Alor Setar, Hospital Baling, Bayan Lepas, Sitiawan, Melaka}. Semasa tempoh MTL, kawasan di bahagian pantai timur lebih banyak menerima hujan berbanding dengan kawasan di bahagian pantai barat (www.met.gov.my). Ini adalah kesan daripada luruan monsun yang diterima dalam tempoh tersebut yang membawa cuaca lembap di kawasan pantai timur. Perbezaan kadar penerimaan jumlah hujan ini menyumbang kepada hasil keputusan kelompok yang diperoleh.

Hasil keputusan kelompok menggunakan ukuran COR pula terdiri daripada kelompok {Kota Bharu, Hospital Dungun}, {Kuantan, Mersing}, {Alor Setar, Hospital Baling} dan {Bayan Lepas, Ipoh, Hospital Tapah, Sitiawan, Subang, Melaka}. Manakala ED mengenal pasti stesen Kuantan dan Mersing sebagai objek terpencil. Stesen selebihnya dikelompokkan bersama iaitu {Alor Setar, Hospital Baling, Ipoh, Hospital Tapah, Sitiawan, Subang, Melaka}.

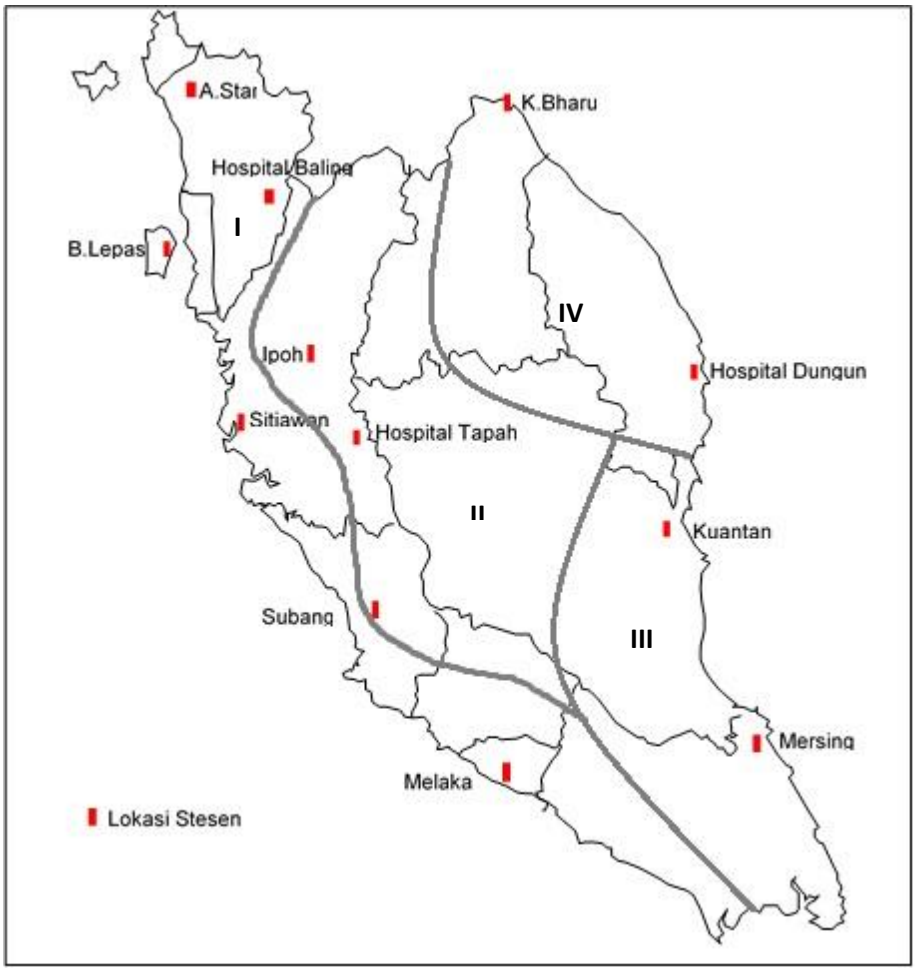
Ringkasan keputusan analisis dendogram bagi setiap ukuran jarak ketaksamaan mengikut kelompok dalam peratusan diberikan dalam Jadual 4.4. Manakala, pembahagian sempadan hasil pengelompokan menggunakan sukatan ketaksamaan CID dan IP, iaitu hasil kelompok terbaik (rujuk Jadual 4.2) untuk data musiman MTL, diilustrasikan dalam Rajah 4.5.



Rajah 4.4 Dendogram untuk 12 stesen bagi analisis data siri masa musiman MTL dengan sukatan ketaksamaan A) Jarak *Euclidean* B) *Complexity-invariant* C) *Integrated Periodogram-based* D) *Correlation-based*.

Jadual 4.4 Peratusan (%) bilangan stesen mengikut kelompok dengan sukatan ketaksamaan berbeza bagi analisis data MTL.

Kelompok	Jarak Sukatan Ketaksamaan			
	ED	CID	IP	COR
#1	66.70%	41.70%	41.70%	50.00%
#2	16.70%	25.00%	25.00%	16.70%
#3	8.30%	16.70%	16.70%	16.70%
#4	8.30%	16.70%	16.70%	16.70%



Rajah 4.5 Peta pengelompokan data hujan Semenanjung Malaysia menggunakan sukatan CID atau IP terhadap data musiman, MTL.

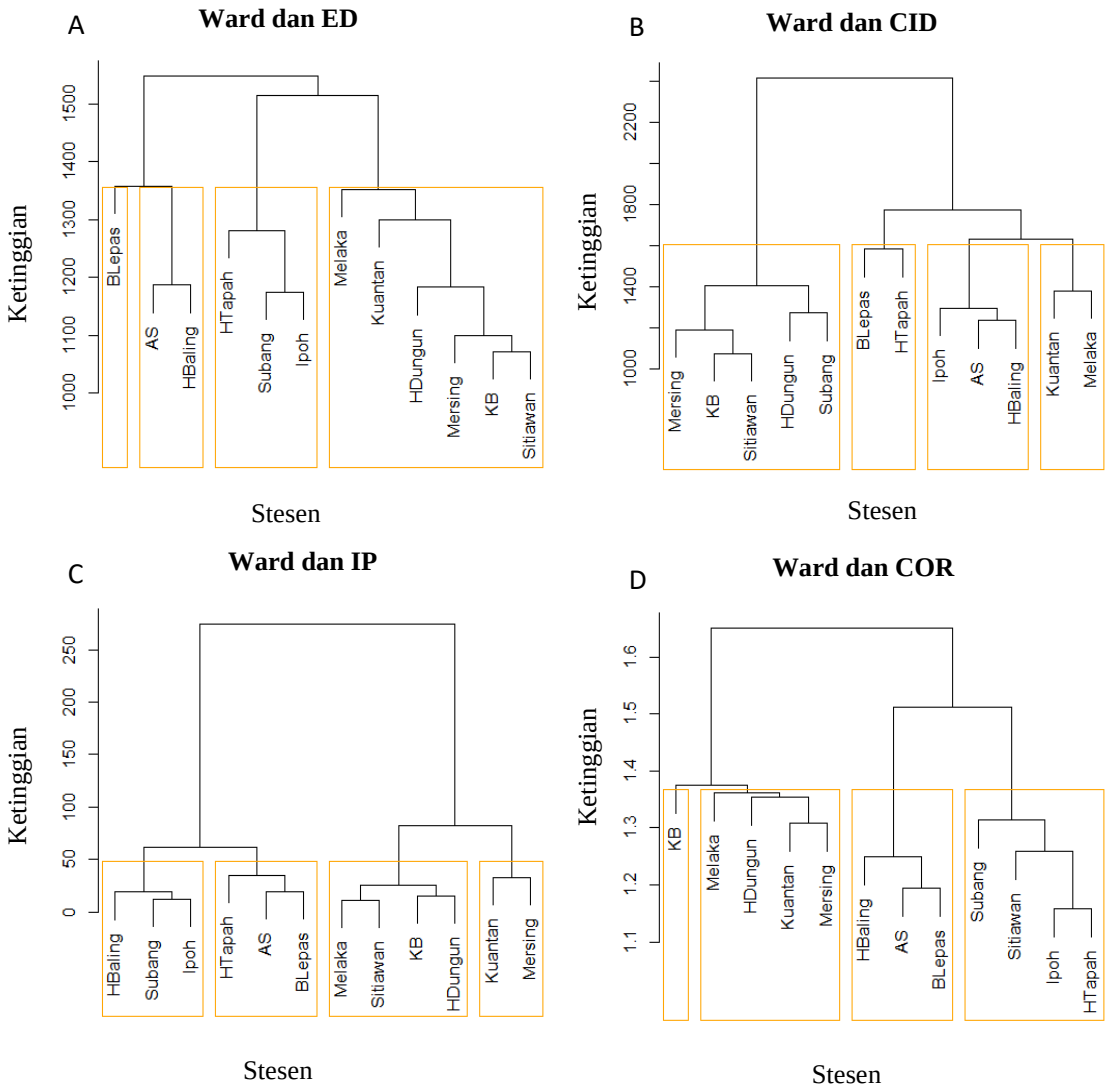
4.5 ANALISIS DENDOGRAM: DATA MUSIMAN MBD

Berdasarkan dendogram dalam Rajah 4.6, diperhatikan hasil analisis pengelompokan yang dilakukan ke atas siri masa MBD agak berbeza dengan hasil yang diperoleh dari analisis data keseluruhan dan MTL. Melalui analisis pengelompokan siri masa bagi data keseluruhan dan MTL diperhatikan faktor geografi dan lokasi stesen berkemungkinan turut memainkan peranan dalam kelompok-kelompok yang terhasil. Namun, menggunakan siri masa MBD, atribut hujan yang diterima dalam tempoh musim tersebut di setiap stesen lebih dominan dalam menentukan kelompok-kelompok yang terhasil. Pada amnya, semasa tempoh MBD, pengurangan curahan hujan berlaku di seluruh negara terutamanya di kawasan pantai barat Semenanjung (www.met.gov.my). Ciri-ciri ini dipercayai menyumbang kepada hasil analisis kelompok dalam tempoh ini.

Sukatan ketaksamaan COR yang dikenal pasti sebagai ukuran yang paling tepat (rujuk Jadual 4.2) untuk digunakan ke atas data siri masa MBD dalam kajian ini memberikan hasil kelompok seperti berikut; {Alor Setar, Hospital Baling, Bayan Lepas}, {Ipoh, Hospital Tapah, Sitiawan, Subang} {Hospital Dungun, Kuantan, Mersing} dan Kota Bharu dikenal pasti sebagai objek terpencil. Hasil analisis menggunakan IP sebagai sukatan ketaksamaan pula memberikan kelompok yang berikut; {Alor Setar, Bayan Lepas, Hospital Tapah}, {Hospital Baling, Ipoh, Subang}, {Sitiawan, Melaka, Kota Bharu, Hospital Dungun} dan {Kuantan, Mersing}.

Melalui penilaian sukatan ketaksamaan bagi analisis pengelompokan dalam Jadual 4.2, dikenal pasti ED dan CID sebagai ukuran yang kurang sesuai dalam menentukan kelompok bagi siri masa MBD. Hasil analisis menggunakan sukatan ketaksamaan ED seperti berikut; {Alor Setar, Hospital Baling}, {Ipoh, Hospital Tapah, Subang}, {Sitiawan, Melaka, Kota Bharu, Hospital Dungun, Kuantan, Mersing} dan Bayan Lepas dikenal pasti sebagai objek tepencil. Kelompok dari sukatan ketaksamaan CID pula adalah seperti berikut; {Alor Setar, Hospital Baling, Ipoh}, {Bayan Lepas, Hospital Tapah}, {Sitiawan, Subang, Kota Bharu, Hospital Dungun, Mersing} dan {Melaka, Kuantan}.

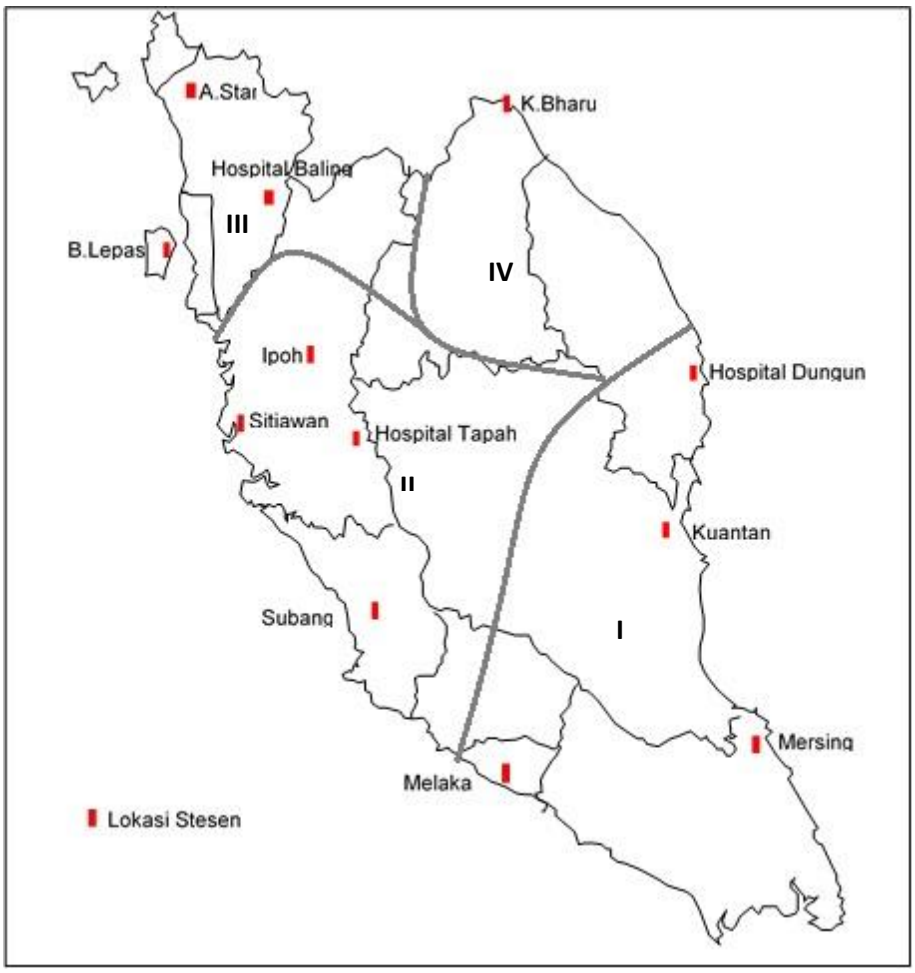
Ringkasan keputusan analisis dendogram bagi setiap ukuran jarak ketaksamaan mengikut kelompok dalam peratusan diberikan dalam Jadual 4.5. Manakala, pembahagian sempadan hasil pengelompokan menggunakan sukatan ketaksamaan COR, iaitu hasil kelompok terbaik (rujuk Jadual 4.2) untuk data musiman MBD, diilustrasikan dalam Rajah 4.7.



Rajah 4.6. Dendogram untuk 12 stesen bagi analisis data siri masa musiman MBD dengan sukatan ketaksamaan A) Jarak *Euclidean* B) *Complexity-invariant* C) *Integrated Periodogram-based* D) *Correlation-based*.

Jadual 4.5 Peratusan (%) bilangan stesen mengikut kelompok dengan sukatan ketaksamaan berbeza bagi analisis data MBD.

Kelompok	**Jarak Sukatan Ketaksamaan**			
	ED	**CID**	**IP**	**COR**
#1	50.00%	41.70%	33.30%	33.30%
#2	25.00%	25.00%	25.00%	33.30%
#3	16.70%	16.70%	25.00%	25.00%
#4	8.30%	16.70%	16.70%	8.30%



Rajah 4.7 Peta pengelompokan data hujan Semenanjung Malaysia menggunakan sukatan COR terhadap data musiman, MBD.

5.0 KESIMPULAN

Pengelompokan siri masa merupakan salah satu kaedah yang boleh digunakan dalam mengkaji corak taburan hujan di Malaysia. Pada amnya, kajian yang dijalankan ini bertujuan mengenal pasti kaedah pengelompokan dan sukatan ketaksamaan yang paling sesuai untuk analisis pengelompokan siri masa data hujan Semenanjung Malaysia. Hasil kajian dapat memberi maklumat berguna dan bermanfaat kepada pelbagai sektor terutamanya berkaitan aspek hidrologi, pengurusan sumber air negara dan pengurusan bencana.

Melalui kajian pengelompokan siri masa data hujan ini, didapati bilangan kelompok yang paling optimum untuk Semenanjung Malaysia ialah empat kelompok. Kelompok-kelompok yang terhasil jelas dapat dibezakan kepada beberapa zon klimatologi yang homogen terutamaya untuk kawasan utara barat laut dan juga pantai timur. Faktor geografi, kedudukan

stesen secara semulajadi dan juga kadar penerimaan hujan sangat mempengaruhi kelompok-kelompok yang terhasil.

Melalui kajian ini juga, didapati bahawa sukatan ketaksamaan IP adalah paling sesuai untuk digunakan ke atas data hujan Semenanjung Malaysia secara keseluruhannya. Manakala bagi data musiman MTL, sukatan ketaksamaan IP atau CID adalah sesuai. Sukatan ketaksamaan COR pula merupakan sukatan paling sesuai untuk digunakan ke atas data hujan musiman MBD. Walau bagaimanapun, perkara utama yang perlu dipertimbangkan dalam kes kajian ini ialah walaupun pengelompokan hierarki Ward dengan sukatan ketaksamaan IP dapat memberikan prestasi yang baik apabila digunakan ke atas data hujan secara keseluruhan dan semasa MTL, ianya tidak membawa maksud bahawa gabungan ini akan bekerja dengan baik secara umum. Hal yang sama perlu dipertimbangkan untuk penggunaan sukatan ketaksamaan CID terhadap data siri masa MTL dan COR terhadap data siri masa MBD.

Melalui keputusan kajian yang diperoleh, bagaimanapun dapat ditunjukkan bahawa adalah satu keperluan untuk meneroka dan melakukan kajian lanjut ke atas topik ini. Pelbagai kaedah pengelompokan dan sukatan ketaksamaan berbeza dengan mengambil kira corak iklim musiman dan juga menggunakan jumlah stesen yang lebih banyak boleh dijalankan di masa hadapan.

RUJUKAN

Acock, A.C. 2005. Working with missing values. *Journal of Marriage and Family* 67: 1012 – 1028.

Bartok, J., Habala, O., Bednar, P., Gazak, M. & Hluchy, L. 2010. Data mining and integration for predicting significant meteorological phenomena. *ICCS* 1(1): 37 – 46.

Bothale, Rajashree & Katpatal, Yashwant. 2014. Spatial and statistical clustering based regionalization of precipitation and trend identification in Prantha catchment, India. *International Journal of Innovative Research in Science, Engineering and Technology* 3(5): 12557-12567.

Cretat, J., Richard, Y., Pohl, B., Rouault, M., Reason, C. & Fauchereau, N. 2010. Recurrent daily rainfall patterns over South Africa and associated dynamics during the core of the austral summer. *International Journal of Climatology* 32(2): 261-273.

Das, R., Bhattacharyya, D.K. & Kalita, J.K. 2007. An effective dissimilarity measure for clustering gene expression time series data. Fourth Biotechnology and Bioinformatics Symposium (BIOT 07). Colorado Springs, CO.

De Gaetano, A.T. 2001. Spatial grouping of United States climates stations using a hybrid clustering approach. *International Journal of Climatology* 21: 791-807.

Ding, Hui, Trajcevski, G., Scheuermann, P., Wang, Xiaoyue & Keogh, E. 2008. Querying and mining of time series data: Experimental comparison of representations and distance measures. *Proceedings of the VLDB Endowment* 1(2): 1542-1552.

Falcone, J.L. & Albuquerque, P. 2004. A correlation-based distance. Short Technical Report. Computer Science Department, University of Geneva.

Faloutsos, C., Ranganathan, M. & Manolopoulos, Y. 1994. Fast subsequence matching in time-series databases. *ACM SIGMOD Record* 23(2): 419-429.

Halkidi, M., Batistakis, Y. & Vazirgiannis, M. 2001. On clustering validation techniques. *Journal of Intelligent Information Systems* 17(2/3): 107-145.

Han, J. & Kamber, M. 2001. *Data Mining: Concepts and Techniques*. San Francisco: Morgan Kaufmann.

Hong, Tianzhen & Jiang, Yi. 1995. Stochastic weather model for building HVAC systems. *Building & Environment* 30(4): 521-532.

Jabatan Meteorologi Malaysia, www.met.gov.my.

Jamaludin Suhaila, Sayang Mohd Deni, Wan Zawiah Wan Zin & Abdul Aziz Jemain. 2010. Trends in Peninsular Malaysia rainfall data during the Southwest Monsoon and Northeast Monsoon seasons: 1975 – 2004. *Sains Malaysiana* 39(4): 533-542.

Kaiser, J. 2014. Dealing with missing values in data. *Journal of Systems Integration* 5(1): 42-51.

Kaufman, L. & Rousseuw, P.J. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley.

Kavitha, V. & Punithavalli, M. 2010. Clustering time series data stream - A literature survey. *International Journal of Computer Science and Information Security* 8(1).

Kohail, S.N. & El-Halees, A.M. 2011. Implementation of data mining techniques for meteorological data analysis (A case study for Gaza Strip). *International Journal of Information and Communication Technology Research* 1(3): 96-100.

Mahmut Firat, Fatih Dikbas, Koc, A.C. & Mahmud Gungor: Missing data analysis and homogeneity test for Turkish precipitation series. *Journal of the Indian Academy of Sciences* 35(6): 707–720.

Mohammadi, K., Eslami, R.H. & Kahawita, R. 2006. Parameter estimation of an ARMA model for river flow forecasting using goal programming. *Journal of Hydrology* 331: 293-299.

Munoz-Diaz, D. & Rodrigo, F.S. 2004. Spatio-temporal patterns of seasonal rainfall in Spain (1912-2000) using cluster and principal component analysis: Comparison. *Annales Geophysicae* 22: 1435-1448.

Norlee Hainie Ahmad & Sayang Mohd Deni. 2013. Homogeneity test on daily rainfall series for Malaysia. *MATEMATIKA* 29: 141-150.

Norlee Hainie Ahmad, I R Othman & Sayang Mohd Deni. 2013. Hierarchical cluster approach for regionalization of Peninsular Malaysia based on the precipitation amount. *Journal of Physics* 423: 12-18.

Montero, P. & Vilar, J.A. 2014. TSclust: An R package for time series clustering. *Journal of Statistical Software* 62(1): 1-43.

Potocnik, P., Berlec, T., Starbek, M. & Govekar, E. 2011. SOM-Based clustering and optimization of production part I. *IFIP AICT* 363: 21-30.

Prasanna, K.A.V.L. & Kumar, V. 2012. Performance evaluation of multiviewpoint-based similarity measure for data clustering. *Journal of Global Research in Computer Science* 3(11): 21-26.

Ramos, M.C. 2001. Divisive and hierarchical clustering techniques to analyse variability of rainfall distribution patterns in a Mediterranean region. *Journal of the Atmospheric Sciences* 57: 123-138.

Sangeeta, R. & Geeta, S. 2012. Recent techniques of clustering of time series data: A survey. *International Journal of Computer Applications* 52(15): 1-9.

Sanwlani, M. & Prof. Vijayalakshmi, M. 2013. Forecasting sales through time series clustering. *International Journal of Data Mining & Knowledge Management Process* 3(1): 39-56.

Soltani, S. & Modarres, R. 2006. Classification of spatio-temporal pattern of rainfall in Iran using a hierarchical and divisive cluster analysis. *Journal of Spatial Hydrology* 6(2): 1-12.

Tennant, W.J & Hewitson, B.C. 2002. Intra-seasonal rainfall characteristics and their importance to the seasonal prediction problem. *International Journal of Climatology* 22(9): 1033-1048.

Wilks, D.S. 2006. *Statistical Methods in the Atmospheric Sciences*. Edisi ke-2. Amsterdam: Elsevier.

WMO. 2011. *WMO Strategic Plan 2012 – 2015*. Geneva: World Meteorological Organization.

MALAYSIAN METEOROLOGICAL DEPARTMENT
JALAN SULTAN
46667 PETALING JAYA
SELANGOR DARUL EHSAN
Tel: 603-79678000
Fax: 603-79550964
www.met.gov.my

ISBN 978-967-5676-95-6